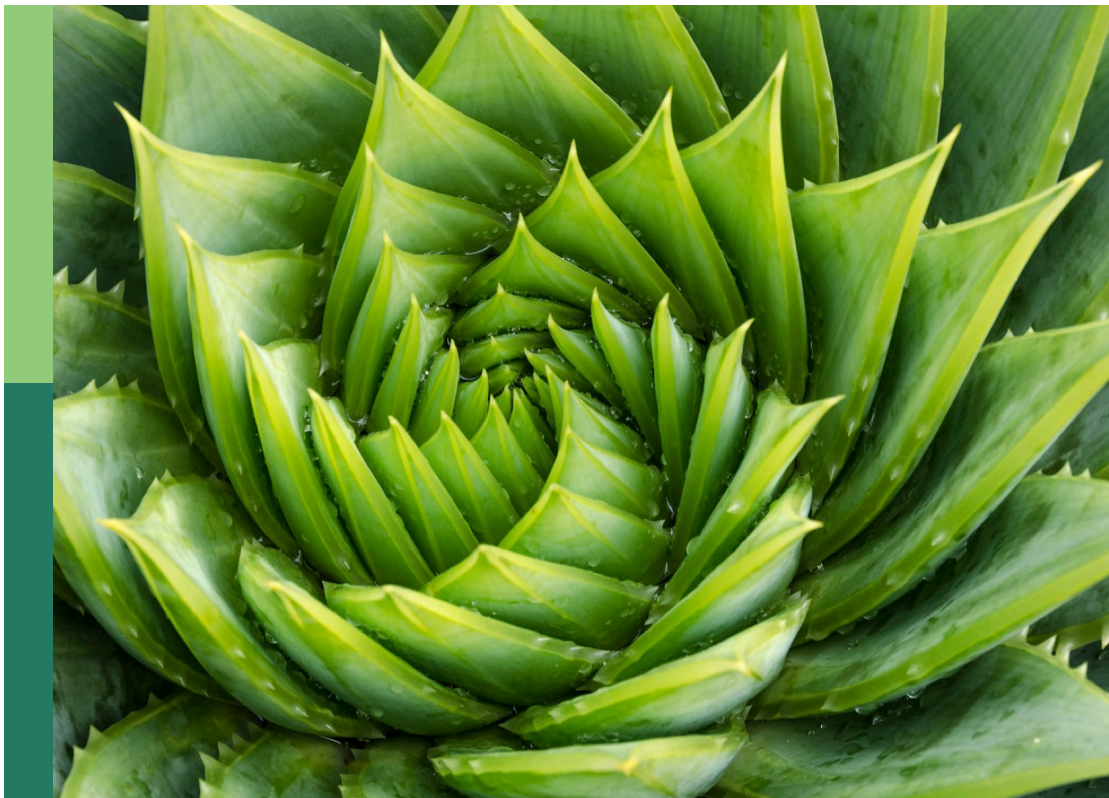
frontiers | Research Topics

Marker-assisted selection (MAS) in crop plants, volume II

Edited by
Ting Peng, Baohua Wang, Muhammad Kashif Riaz Khan and Peng Chee

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-5035-9
DOI 10.3389/978-2-8325-5035-9

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Marker-assisted selection (MAS) in crop plants, volume II

Topic editors

Ting Peng — Henan Agricultural University, China

Baohua Wang — Nantong University, China

Muhammad Kashif Riaz Khan — Nuclear Institute for Agriculture and Biology, Pakistan

Peng Chee — The University of Georgia, Tifton Campus, United States

Citation

Peng, T., Wang, B., Khan, M. K. R., Chee, P., eds. (2024). *Marker-assisted selection (MAS) in crop plants, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5035-9

Table of contents

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 25 January 2023
DOI 10.3389/fpls.2023.1113059

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Wànkuí Gong,
Institute of Cotton Research (CAAS), China
Hantao Wang,
Institute of Cotton Research (CAAS), China

*CORRESPONDENCE
Miao Li
limiao2003@sina.com

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 01 December 2022
ACCEPTED 09 January 2023
PUBLISHED 25 January 2023

CITATION
Jia X, Wang S, Zhao H, Zhu J, Li M and Wang G (2023) QTL mapping and BSA-seq map a major QTL for the node of the first fruiting branch in cotton.
Front. Plant Sci. 14:1113059.
doi: 10.3389/fpls.2023.1113059

QTL mapping and BSA-seq map a major QTL for the node of the first fruiting branch in cotton

Xiaoyun Jia, Shijie Wang, Hongxia Zhao, Jijie Zhu, Miao Li* and Guoyin Wang

Institution of Cereal and Oil Crops, Hebei Academy of Agriculture and Forestry Sciences/Hebei Laboratory of Crop Genetics and Breeding/Hebei Key Laboratory of Crop Cultivation Physiology and Green Production, Shijiazhuang, China

Understanding the genetic basis of the node of the first fruiting branch (NFFB) improves early-maturity cotton breeding. Here we report QTL mapping on 200 F_2 plants and derivative $F_{2:3}$ and $F_{2:4}$ populations by genotyping by sequencing (GBS). BC_1F_2 population was constructed by backcrossing one $F_{2:4}$ line with the maternal parent JF914 and used for BSA-seq for further QTL mapping. A total of 1,305,642 SNPs were developed between the parents by GBS, and 2,907,790 SNPs were detected by BSA-seq. A high-density genetic map was constructed containing 11,488 SNPs and spanning 4,202.12 cM in length. A total of 13 QTL were mapped in the 3 tested populations. JF914 conferred favorable alleles for 11 QTL, and JF173 conferred favorable alleles for the other 2 QTL. Two stable QTL were repeatedly mapped in $F_{2:3}$ and $F_{2:4,}$ including *qNFFB-D3-1* and *qNFFB-D6-1*. Only *qNFFB-D3-1* contributed more than 10% of the phenotypic variation. This QTL covered about 24.7 Mb (17,130,008–41,839,226 bp) on chromosome D3. Two regions on D3 (41,779,195–41,836,120 bp, 41,836,768–41,872,287 bp) were found by BSA-seq and covered about 92.4 Kb. This 92.4 Kb region overlapped with the stable QTL *qNFFB-D3-1* and contained 8 annotated genes. By qRT-PCR, *Ghir_D03G012430* showed a lower expression level from the 1- to 2-leaf stage and a higher expression level from the 3- to 6-leaf stage in the buds of JF173 than that of JF914. *Ghir_D03G012390* reached the highest level at the 3- and 5-leaf stages in the buds of JF173 and JF914, respectively. As JF173 has lower NFFB and more early maturity than JF914, these two genes might be important in cell division and differentiation during NFFB formation in the seedling stage. The results of this study will facilitate a better understanding of the genetic basis of NFFB and benefit cotton molecular breeding for improving earliness traits.

KEYWORDS

cotton earliness, node of the first fruiting branch, QTL mapping, BSA-seq, candidate gene

Introduction

Upland cotton (*Gossypium hirsutum* L. AADD, 2n=52) is the most important fiber crop in the world, accounting for more than 90% of global cotton production (Chen et al., 2007; Ma et al., 2021). Cottonseed is also a good source of edible oil and vegetable protein (Zhang et al., 2015). Thus, upland cotton has significant value in dealing with the increasing human population. Earliness is one of the vital breeding goals to meet the needs of mechanism practice, especially during cotton harvesting (Jia et al., 2016; Li et al., 2021). Besides, early-maturity cotton, also known as short-season cotton, has many advantages in inter-cropping between cereal crops and cotton to increase land utilization efficiency in China (Cheng et al., 2021; Zhao et al., 2022). Earliness is a typical characteristic of early-maturity cotton. As yield and fiber quality have dominated cotton breeding for decades, little attention has been paid to earliness.

In terms of plant development, cotton earliness is described as flowering time (FT), whole growth period (WGP), and flowering-to-boll opening period (FBP) (Richmond and Radwan, 1962; Li et al., 2020). Plant height (PH), node of the first fruiting branch (NFFB), and height of NFFB (HNFFB) are also important indexes for earliness (Godoy and Palomo, 1999; Jia et al., 2016). NFFB is the most reliable index in terms of indicating cotton earliness, which has better consistency among environments, and significantly positively correlates with FT, WGP, PH, and HNFFB (Guo et al., 2008; Su et al., 2016; Zhang et al., 2021). All six traits mentioned above have relatively high broad-sense heritabilities and significant environmental influences (Jia et al., 2016; Li et al., 2020; Li et al., 2021).

Several studies for cotton earliness genetic detection through QTL mapping and GWAS analysis have been published (Li et al., 2020). Guo et al. (2008); Guo et al., 2009 mapped QTL for NFFB in two F_2 populations and used the results to measure flowering time. Li et al. (2013) mapped 54 QTL for cotton earliness in two F_2 and their $F_{2:3}$ populations, and a common QTL for the budding period could explain 12.6% of the phenotypic variation. Benefiting from high-throughput sequencing techniques and high-quality genome sequences of TM-1 and NDM8, the efficiency and accuracy of QTL mapping and GWAS analysis have been significantly improved (Li et al., 2015; Zhang et al., 2015; Hu et al., 2019; Wang et al., 2019; Ma et al., 2021). Jia et al. (2016) constructed a high-density genetic map containing 6295 SNP and 139 SSR markers for a RIL population by RAD-seq, mapped 247 QTL for cotton earliness in six consecutive years, and found an extremely prominent chromosome region on D3 with six stable major QTL. Li et al. (2017) constructed a SNP-based genetic map for an F_2 population by GBS, mapped 47 QTL for cotton earliness, and found a major region on D3 overlapping with the results of Jia et al. (2016). Su et al. (2016) developed 81,675 SNPs in 355 upland cotton accessions; 13 significant associations between SNP and earliness traits were found by GWAS, a major locus and a candidate gene were also mapped on D3. Li et al. (2021) re-sequenced 436 cotton accessions and developed 10,118,884 SNPs and 864,132 InDels; 307 significant SNPs were found for cotton earliness by GWAS, including 43 SNPs in a 3.7 Mb region on D3 overlapping with previous results. The reports mentioned above imply the significant role of chromosome D3 in controlling cotton earliness, which has been emphasized again by Ma et al. (2018) and Zhang et al. (2021). Besides, Li et al. (2018) developed 49,650 SNPs in 169 upland cotton accessions by CottonSNP80K array; 29 significant SNPs and two candidate genes were found for cotton earliness. However, QTL fine mapping for cotton earliness, especially NFFB, has rarely been reported until now, and the genetic basis under earliness traits is still unclear.

This study used a nationally certified variety, Jifeng914 (JF914), with about 120 d WGP and 8 NFFB as the maternal parent, an early maturity inbred line Jifeng173 (JF173) with about 108 d WGP and 5 NFFB was used as the paternal parent. QTL mapping was conducted based on a high-density genetic map for an F_2 population. The BC_1F_2 population was constructed and used for QTL mapping by BSA-seq. One stable QTL for NFFB spanning 24.7 Mb was shortened to 92.4 Kb. Eight genes were annotated in this core region and 2 genes with different expression patterns in the buds of JF173 and JF914 might be the candidates.

Materials and methods

Experimental materials and phenotypic trait

Jifeng 914 (JF914) (a larger phenotype cultivar with about 120 d WGP and 8 NFFB) was crossed with Jifeng 173 (JF173) (a smaller phenotype inbred line with about 108 WGP and 5 NFFB). An F_2 population containing 417 plants was developed at Shijiazhuang in 2019; 200 plants from the F_2 were randomly selected and continuously self-pollinated to $F_{2:3}$ and $F_{2:4}$ generations. The $F_{2:3}$ and $F_{2:4}$ populations were planted with two replicates in a completely randomized block design at Shijiazhuang in 2020 and 2021. One $F_{2:4}$ line with low NFFB and a similar phenotype to JF914 was chosen and backcrossed with JF914 in 2021. And 23 BC_1 plants were self-pollinated at Hainan in the winter of 2021 to construct the BC_1F_2 population containing 561 plants, which was planted in 2022 at Shijiazhuang. The materials were planted in single lines (5 m long and 70 cm between adjacent lines), and conventional field management was carried out.

The node of the first fruiting branch (NFFB) was tested. Every plant in the F_2 and BC_1F_2 was measured. Ten plants in the middle of each line were measured in the $F_{2:3}$ and $F_{2:4}$ populations. Excel 2010 and SPSS 17 were used for data analysis.

DNA sequencing

Genomic DNA was extracted by the CTAB method (Paterson et al., 1993). The genotyping-by-sequencing (GBS) method was applied for the F_2 plants as detailed by Zhang et al. (2016); Li et al. (2017), and Zhou et al. (2016). Briefly, DNA was incubated at 37°C with *Mse* I (New England Biolabs, NEB), T4 DNA ligase (NEB), ATP 9NEB, and *Mse* Y adapter N containing barcodes. *Hae* III and *Rco*R I (NEB) were added into the *Mse*I digestions to further digest the fragments at 37°C. Fragments of 397–420 bp were purified and paired-end 150-bp sequenced on the Illumina HiSeq™ platform. High-quality reads were filtered based on (1) removing reads with ≥ 10% unidentified nucleotides (N); (2) removing reads with > 50% based on having Phred quality < 5; (3) removing reads with 10 nt

aligned to the adapter, allowing ≤ 10% mismatches; and (4) removing reads containing *Hae* III or *Eco*R I.

For BSA-seq, 30 high NFFB plants and 30 low NFFB plants were selected from the BC_1F_2 population; the DNA of each plant was extracted and mixed to construct two DNA pools (high and low). Four samples were subjected to re-sequencing, including JF914, JF173, and high and low DNA pools. The GenoBaits DNA-seq Library Prep kit was used for library construction. First, 4 μl of GenoBaits End Repair Buffer and 2.7 μl of GenoBaits End Repair Enzyme were added to 200 ng DNA and incubated for 20min at 37°C and 20min at 72°C. Second, 2 μl of GenoBaits Ultra DNA ligase, 8 μl of GenoBaits Ultra DNA ligase Buffer, and 2 μl of GenoBaits Adapter were added and incubated for 30 min at 22°C. Third, fragments were purified by adding 48 μl of Beckman AMPure XP Beads. Fragments of 200–300 bp were reserved and sequenced on the MGI-2000/MGI-T7 planform. High-quality reads were filtered based on (1) removing the adaptor; (2) removing reads with >10% N; and (3) removing reads with >40% low-quality bases (Q ≤ 20).

The BWA software was used to align the clean reads against the reference genome of TM-1 (Wang et al., 2019). The GATK software was used for variation calling (Mckenna et al., 2010). SnpEff and ANNOVAR software were used for annotation (Wang et al., 2010; Cingolani et al., 2014).

QTL mapping

Polymorphic markers developed from the F_2 population were classified into eight segregation patterns (aa×bb, ab×cc, ab×cd, cc×ab, ef×eg, hk×hk, lm×ll, nn×np), and the aa×bb pattern SNPs were chosen to construct the genetic map. SNPs with segregation distortion ($p<0.001$) or integrity (<40%) or in the same reads or abnormal bases were filtered. SNP markers were sorted into 26 chromosomes according to their physical position on the reference genome. And then, the genetic map was constructed chromosome by chromosome using JoinMap 4.0 with a LOD score threshold of 6.0–20.0. The ICIM method in the QTL IciMapping software was used to detect QTL (Li et al., 2007). Parameters were set as 1 cM per step, PIN=0.001, and the LOD score was determined by a 1000 permutation test.

The Δ(SNP-index) and ED (Euclidean distance) methods were used to analyze the candidate region between the pools. The parameters of SNP-index and Δ(SNP-index) were calculated as follows: SNP-index(high) = Mhigh/(Mhigh + Phigh), SNP-index (low) = Mlow/(Mlow + Plow), and Δ(SNP-index) = SNP-index (low) – SNP-index(high). The M and P parameters represent the sequencing depth in JF914 and JF173, respectively. The parameters of ED were calculated as follows:

$$ED=\sqrt{(Ahigh\text{-}Alow)^2+(Thigh\text{-}Tlow)^2+(Chigh\text{-}Clow)^2+(Ghigh\text{-}Glow)^2}$$

A, T, C, and G are the four base types. Ahigh, Thigh, Chigh, and Ghigh are the frequency of relevant bases in the high pool. Alow, Tlow, Clow, and Glow are the frequency of relevant bases in the low pool. The ED^4 was used to eliminate background noise. The median +3SD was used as the threshold.

qRT-PCR

For JF914 and JF173, total RNA was extracted from the bud and leaf at 1- to 6-leaf stages using a Plant RNA Purification Kit (Tiangen, Beijing, China). First-strand cDNA was reverse transcribed from 1 μg total RNA using a FastKing gDNA Dispelling RT SuperMix Kit (Tiangen, Beijing, China). qRT-PCR was carried out with the SYBR Premix Ex Taq (TAKARA, Japan) on a LightCycler480 instrument (Rotkreuz, Switzerland).

Results

Phenotypic variation

The NFFB of JF914 (7.8–8.5) is significantly bigger than that of JF173 (5.1–5.5). The maximum and minimum values of NFFB in the F_2, $F_{2:3}$, and $F_{2:4}$ populations reveal transgressive segregation (Table 1). The mean value of NFFB in the segregation populations lies within the range of the two parents. Based on the absolute values of skewness and kurtosis, NFFB showed an approximately normal distribution.

Sequence data and quality

A total of 416 G sequence data was obtained by genotyping by GBS, with an average of 25.91 G and 9× depth in the parents, 1.82 G and 0.7× depth in the F_2 plants. The Q30 score reached 95.68%. And 99.63% of the F_2 sequence data was successfully mapped to the reference genome, with an average coverage rate of 14.81% (Additional file 1).

A total of 318.81 G sequence data was obtained by re-sequencing for the four samples (Table 2). The sequence depths of the pools reached 31×, and Q30 scores are larger than 90%. More than 88% of the reference genome was covered.

TABLE 1 The statistics of NFFB in the parents, F_1, and segregated populations.

Trait	JF914	JF173	F_2					$F_{2:3}$					$F_{2:4}$				
			Max	Min	Mean	Skew	Kurt	Max	Min	Mean	Skew	Kurt	Max	Min	Mean	Skew	Kurt
NFFB (cm)	7.8-8.5**	5.1-5.5	9.0	5.0	6.6	-0.4	0.3	8.1	5.5	6.6	-0.2	0.3	8.7	5.1	6.3	0.1	0.3

NFFB, the node of the first fruiting branch; Max, maximum; Min, minimum; Skew, skewness; Kurt, kurtosis; **, $p<0.01$.

TABLE 2 Sequence data of the parents and pools.

Sample	Raw Bases (bp)	Clean Bases (bp)	Q20 (%)	Q30 (%)	Align rate (%)	Average depth (×)	Coverage (%)
high	100,641,873,900	99,983,581,254	96.58	90.50	78.73	31.99	88.17
low	143,699,754,900	142,505,777,730	96.66	91.09	79.36	34.63	88.40
JF914	26,169,828,600	26,013,215,006	96.4	89.87	81.46	8.77	86.80
JF173	48,300,015,000	47,983,815,204	96.63	90.63	81.10	15.62	87.67

Genetic map construction

A total of 1,305,642 SNPs were developed between the parents, and 7 SNP types were detected (Table 3). Only the SNP in aa×bb type was used to genotype the F_2 plants and construct a genetic map. A high-density genetic map containing 11,488 SNPs was constructed (Figure 1, Table 4, Additional file 2). The genetic map spanned 4,202.12 cM in length, ranging from 150.74 cM on A3 to 178.90 cM on A9. The SNPs were unevenly distributed on the 26 linkage groups, with only 30 SNPs on A2 and 1 318 SNPs on D5. The quality of the genetic map was analyzed by colinearity analysis, which demonstrated the accurate SNP position on the constructed genetic map (Table 4, Figure 2).

QTL mapping

A total of 13 QTL were mapped and distributed on 11 chromosomes (Table 5). Two stable QTL were mapped, including *qNFFB-D3-1* and *qNFFB-D6-1* mapped in $F_{2:3}$ and $F_{2:4}$. JF914 conferred favorable alleles for 11 QTL, and JF173 conferred for the other two QTL. Only *qNFFB-D3-1* contributed more than 10% of the phenotypic variation. Thus, this QTL might be vital loci regulating NFFB in the tested population.

A total of 2,907,790 SNPs were detected by BSA-seq, including 1,926,811 transition types and 979,643 transversion types (Figure 3). After filtration, 348,074 high-quality SNPs were reserved (Additional file 3). And SNP index was calculated for 337,651 SNPs (Additional file 4). A total of 197 and 99 regions were found through Δ(SNP-index) analysis and ED analysis, respectively (Additional file 5). Thirty-nine regions containing 2310 SNPs on 12 chromosomes were common between the results of Δ(SNP-index) analysis and ED analysis, which were recognized as the candidate regions for NFFB (Additional file 6). Two regions on D3 (41,779,195–41,836,120 bp and 41,836,768–41,872,287 bp) overlapped with the stable QTL *qNFFB-D3-1* (17,130,008–41,839,226 bp). Thus, the 24.7 Mb interval of *qNFFB-D3-1* might be shortened to the 92.4 Kb key interval.

TABLE 3 Parent marker types and the number of SNPs.

Marker type	SNP number
aa×bb	410726
ab×cc	159
cc×ab	47
ef×eg	340
hk×hk	325126
lm×ll	162538
nn×np	406706

Candidate gene analysis

Eight genes were annotated in the 92.4 Kb interval of the stable QTL *qNFFB-D3-1* (D3, 41,779,195–41,836,120 bp and 41,836,768–41,872,287 bp) (Table 6). By qRT-PCR, 2 genes showed significant different and regular expression patterns between the buds of JF914 and JF173 (Figure 4). *Ghir_D03G012430* expressed at a lower level at 1- and 2-leaf stages and increased sharply to a higher expression level at 3- to 6-leaf stages in the bud of JF173 than that of JF914. *Ghir_D03G012390* reached the highest expression level in the buds of JF173 and JF914 at 3- and 5-true leaf stages, respectively. The expression levels of the above mentioned genes in leaves showed no regular patterns. Thus, these 2 genes might be involved in NFFB regulation.

Discussion

As a labor-intensive crop, cotton is increasingly unsuitable for manual planting in China, which raises the very pressing need for whole-process mechanization (Ma et al., 2019). Earliness is a vital trait for the practice of mechanism. Xinjiang is one of the most important cotton-growing regions in the world, accounting for 84.94% of China and ~19% of the world of cotton production (Han et al., 2020). Unstable weather conditions during the cotton planting season may cause heavy losses, especially in northern Xinjiang. Late sowing by planting early-maturity cotton is a useful method to avoid adverse weather in spring (Cheng et al., 2021). Besides, early-maturity cotton can optimize farmland cropping systems by directly planting cotton after wheat harvesting (Li et al., 2020). Thus, to improve efficiency and breed early maturity varieties suitable for mechanical harvesting, there is more need for the genetic detection of cotton earliness. In this study, an F_2 population containing 417 plants was constructed to map QTL for cotton earliness. High-quality and density SNP markers were detected by high-throughput genome sequencing. A high-density genetic map containing 11,488 SNP and spanning 4,202.12 cM was constructed using 200 F_2 plants, which is comparable with the genetic maps used for cotton earliness-related QTL mapping previously reported by Jia et al. (2016) (6434 loci, 4071.98 cM, 137 RILs) and Li et al. (2017) (3978 SNP, 2480 cM, 170 F_2 plants).

The genetic basis of earliness-related traits is complex, involving WGP, FT, FBP, PH, NFFB, and HNFFB, all of which are quantitative traits controlled by large amounts of minor effect genes (Lacape et al.,

10.3389/fpls.2023.1113059

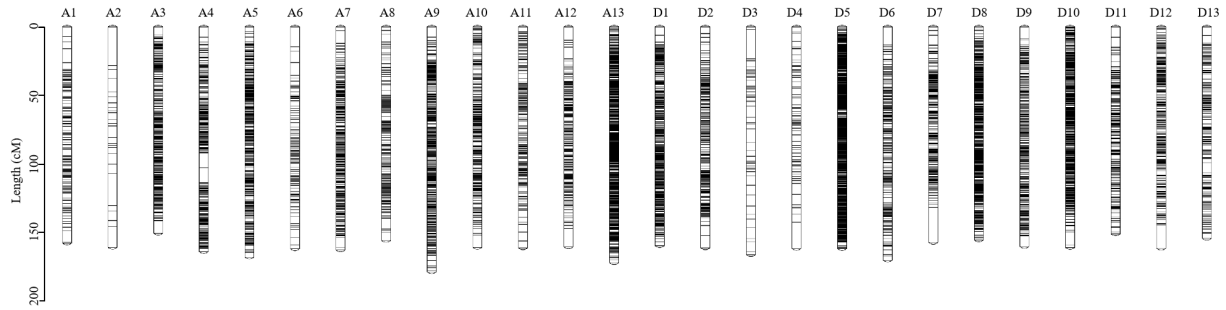


FIGURE 1
Marker distribution of the constructed genetic map.

TABLE 4 Detailed information on the genetic map.

Chr.	No. of Marker	Length (cM)	Average interval (cM)	Largest gap (cM)	Coe. of collinearity
A1	165	157.86	0.96	10.09	-1.00
A2	30	161.10	5.56	28.19	-1.00
A3	487	150.74	0.31	9.25	-1.00
A4	514	164.16	0.32	10.77	-0.99
A5	488	168.17	0.35	3.43	-1.00
A6	152	162.13	1.07	14.63	-1.00
A7	583	162.98	0.28	9.13	-0.90
A8	414	156.02	0.38	7.74	-0.84
A9	624	178.90	0.29	7.61	-1.00
A10	415	161.14	0.39	8.91	-0.99
A11	338	161.48	0.48	6.49	-1.00
A12	270	160.78	0.60	13.61	-1.00
A13	1115	172.27	0.15	2.51	-0.82
At	5595	2117.74	0.74	28.19	–
D1	752	159.88	0.21	6.18	-0.68
D2	311	161.47	0.52	8.92	-1.00
D3	83	166.45	2.03	21.48	-1.00
D4	82	161.89	2.00	19.21	-1.00
D5	1318	162.01	0.12	2.60	-1.00
D6	241	170.41	0.71	13.22	-1.00
D7	359	157.39	0.44	25.66	-1.00
D8	900	155.69	0.17	2.66	-0.97
D9	349	160.44	0.46	8.78	-1.00
D10	743	161.22	0.22	9.77	-0.98
D11	236	151.28	0.64	7.66	-1.00
D12	323	161.79	0.50	17.07	-1.00
D13	196	154.46	0.79	7.64	-1.00
Dt	5893	2084.38	0.61	25.67	–
total	11488	4202.12	0.37	28.19	–

Chr., chromosome; No., number; Coe., coefficient; cM, centi morgan.

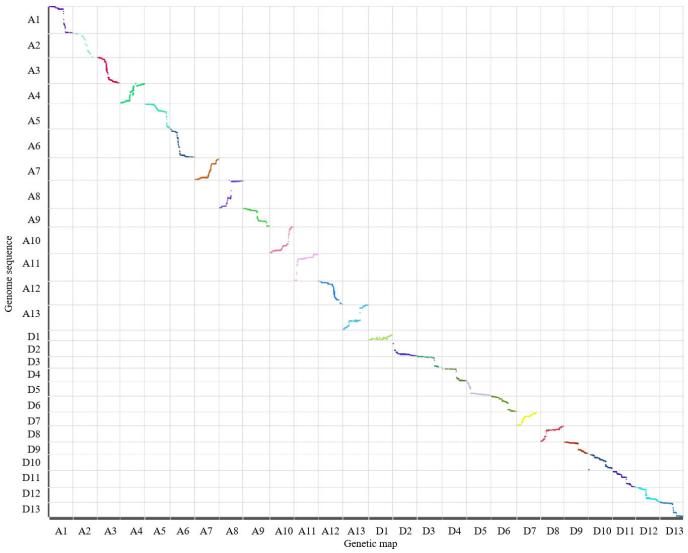


FIGURE 2
Colinearity analysis between the genetic map and reference genome sequence.

2013; Su et al., 2016; Li et al., 2021). The 247 QTL reported by Jia et al. (2016) could explain 0.28–29.37% of the phenotypic variation, and 52 QTL could be detected in at least 2 years. The 47 QTL reported by Li et al. (2017) could explain 3.07–32.57% of the phenotypic variation, and none could be detected repeatedly. The SNPs for earliness traits detected by GWAS could explain 5.36%-15.56% of the phenotypic variation (Su et al., 2016). This study mapped 13 QTL with a 4.74–10.11% phenotypic variation explanation rate for NFFB. Two QTL could be detected in 2 generations, including *qNFFB-D3-1* and *qNFFB-D6-1*, and *qNFFB-D3-1* explained more than 10% of the phenotypic variation. At the same time, it is difficult to dissect the genetic basis under cotton earliness clearly, of the lack of both major and stable QTL (Li et al., 2020).

NFFB is an important index for earliness, such as in cotton (Jia et al., 2016) and pepper (Zhang et al., 2019). And NFFB was considered the most reliable and practical measurement of cotton

TABLE 5 Detailed information of the mapped QTL.

QTL Name	Pop	Pos (cM)	Left marker	Right marker	LOD	PV (%)	Add	Dom
qNFFB-A4-1	F_2	40.00	chr4_75497483	chr4_75488413	2.81	5.03	-0.04	0.39
qNFFB-A7-1	$F_{2:4}$	24.00	chr7_96247777	chr7_92676059	5.13	9.03	-0.14	0.09
qNFFB-A7-2	$F_{2:3}$	28.00	chr7_92674198	chr7_92670233	3.50	6.93	-0.18	-0.01
qNFFB-A11-1	F_2	3.00	chr11_119686364	chr11_119649722	4.03	8.63	-0.05	0.84
qNFFB-D2-1	$F_{2:3}$	71.00	chr15_61609865	chr15_61609728	2.73	5.30	-0.13	-0.07
qNFFB-D3-1	$F_{2:3}$	51.00	chr16_41836768	chr16_17130088	4.16	8.21	-0.18	0.01
	$F_{2:4}$	50.00	chr16_41839226	chr16_41836768	5.62	10.11	-0.17	0.13
qNFFB-D5-1	$F_{2:3}$	9.00	chr18_61260861	chr18_61249526	3.30	6.35	0.00	-0.47
qNFFB-D6-1	$F_{2:3}$	106.00	chr19_12617417	chr19_12617401	3.34	6.52	0.06	-0.24
	$F_{2:4}$	106.00	chr19_12617417	chr19_12617401	3.43	5.75	0.06	-0.18
qNFFB-D7-1	$F_{2:4}$	95.00	chr20_15136585	chr20_14955163	4.35	7.49	-0.16	-0.02
qNFFB-D8-1	F_2	139.00	chr21_5256024	chr21_5235564	3.01	5.87	-0.27	0.00
qNFFB-D10-1	F_2	161.00	chr23_67766849	chr23_67763158	3.51	6.71	-0.31	-0.20
qNFFB-D12-1	F_2	0.00	chr25_62606647	chr25_62552111	2.68	5.06	-0.04	-0.45
qNFFB-D12-2	F_2	138.00	chr25_2807174	chr25_2714956	2.57	4.74	-0.17	-0.26

Pop, population; Pos, position; PV, phenotypic variation; Add, additive effect; Dom, dominance effect; Note: PH, plant height; NFFB, the node of the first fruiting branch; FBP, flowering-to-boll opening period; FT, flowering timing; WGP, whole growth period.

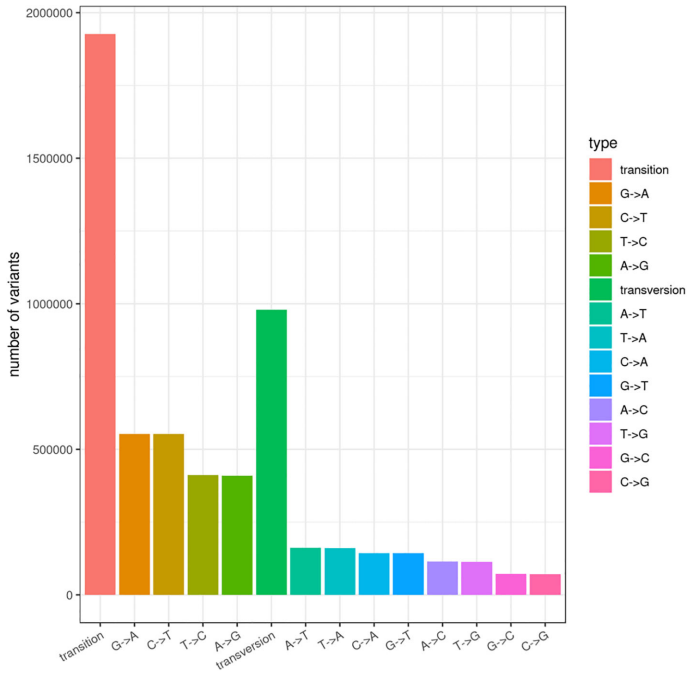


FIGURE 3
Statistic number of each SNP type.

earliness (Ray and Richmond, 1966; Guo et al., 2008). Previously, at least 80 QTL for NFFB were mapped on almost all 26 cotton chromosomes and most of these QTL have tiny genetic effect (Guo et al., 2008; Guo et al., 2009; Li et al., 2012; Jia et al., 2016; Li et al., 2017). As a typical quantitative trait, map a stable major QTL for NFFB is very precious for excavating candidate genes. The chromosome D3 was repeatedly mapped with outstanding QTL: by Jia et al. (2016); Su et al. (2016); Li et al. (2017), and Ma et al. (2018). Thus, it is interesting and hopeful that D3 contains vital genes regulating NFFB. In this study, one stable QTL *qNFFB-D3-1* was mapped in $F_{2:3}$ and $F_{2:4}$ generations and explained 8.21–10.11% of phenotypic variation. The confidence interval of *qNFFB-D3-1* locates between 17.1 to 41.8 Mb, spans a long region of about 24.7 Mb. QTL at this region have been reported repeatedly such as by Jia et al. (2016) (*qNFFB-D3-1*, *qNFFB-D3-2*, *qNFFB-D3-3*, *qNFFB-D3-4*), Li et al. (2017) (*qNFFB-D3-1*), Li et al. (2021) (*rsD03_39122594*), and Zhang et al. (2021) (*qNFFB-Dt3-3*). Candidate genes for cotton earliness in this region were found, such as *GhEMF2* by Jia et al. (2016) and Ma et al. (2020), *Gh_D03G0885* and *Gh_D03G0922* by Li et al. (2017), *Ghir_D03G011310* by Li et al. (2021), and *GhAPL* and *GhHAD5* by Zhang et al. (2021). Other candidate genes for earliness on chromosome D3 were reported, such as *GhCIP1* and *GhUCE* by Ma et al. (2018) and *CotAD_01947* by Su et al. (2016). Thus, it seems likely that *qNFFB-D3-1* contains candidate genes for cotton earliness.

In recent years, BSA-seq has become an efficient method in QTL mapping and functional gene mining and has been widely applied, such as in rice (Takagi et al., 2013; Zhang et al., 2021), tomato (Illa-Berenguer et al., 2015), melon (Hu et al., 2022), *Brassica napus* (Ye et al., 2022), maize (Chen et al., 2021), and cucumber (Lu et al., 2014). In cotton, genes controlling oil content (Liu et al., 2020), virescent

TABLE 6 The eight annotated genes in the 92.4 Kb interval.

Gene ID	Gene Name	Description
Ghir_D03G012380	Bicc1	Protein bicaudal C homolog 1
Ghir_D03G012390	FAM214B	Protein FAM214B
Ghir_D03G012400	At1g54610	Probable serine/threonine-protein kinase At1g54610
Ghir_D03G012410	AGAP005782	ATPase ASNA1 homolog
Ghir_D03G012420	SAE1B-2	SUMO-activating enzyme submit 1B-2
Ghir_D03G012430	pan1	Actin cytoskeleton-regulatory complex protein pan1
Ghir_D03G012440	HSD1	11-beta-hydroxysteroid dehydrogenase 1B
Ghir_D03G012450	RPL7A-2	60S ribosomal protein L7a-2

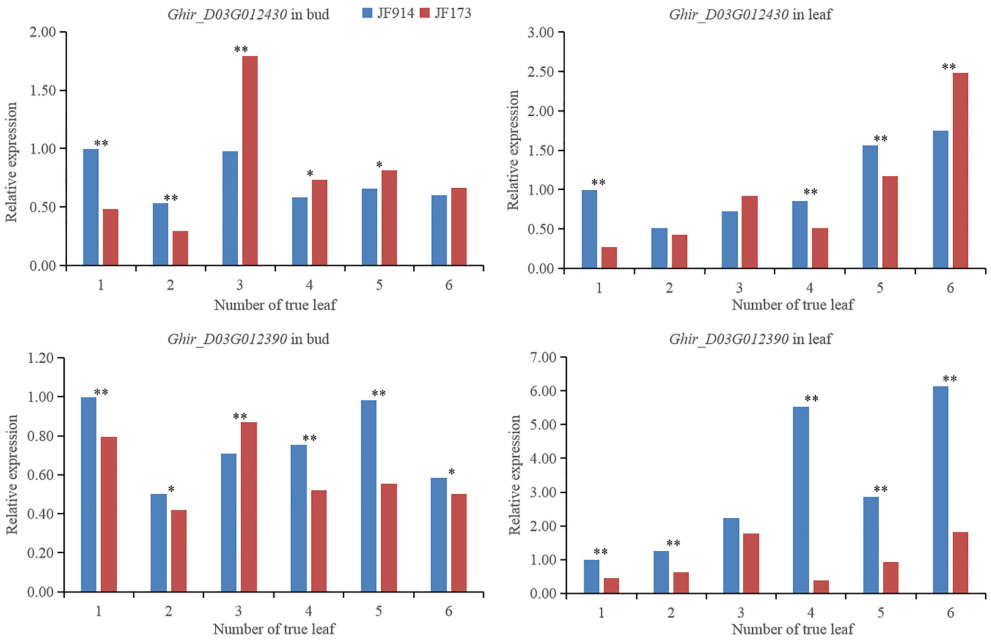


FIGURE 4
Gene expression level in the bud and leaf of JF914 and JF173. *, the difference reached p=0.05 significance level; **, the difference reached p=0.01 significance level.

(Zhu et al., 2017; Gao et al., 2021), nulliplex-branch (Chen et al., 2015; Wen et al., 2021), and NFFB (Zhang et al., 2021) were mapped by BSA-seq. By combining QTL mapping and BSA-seq, QTL can be finely mapped to a very small interval, significantly improving the mining efficiency of vital genes under important quantitative traits (Chen et al., 2022; Hu et al., 2022). In this study, aiming to map candidate genes for NFFB, one line from the $F_{2:4}$ population with low NFFB and similar phenotype to JF914 was used as the maternal parent and backcrossed with JF914. A BC_1F_2 population containing 561 plants was constructed. A total of 60 plants with extremely high (30 plants) or low (30 plants) NFFB from the BC_1F_2 population were selected to construct the high and low pools. And 39 candidate regions were found by Δ(SNP-index) and ED methods. Two regions on D3 (41,779,195–41,836,120 bp, 41,836,768–41,872,287 bp) overlapped with the stable QTL *qNFFB-D3-1* (17,130,008–41,839,226 bp). Thus, the stable QTL *qNFFB-D3-1* spanning 24.7 Mb was shortened to 92.4 Kb key interval, and eight genes were annotated.

By qRT-PCR, *Ghir_D03G012430* was expressed at a lower level at 1- and 2-leaf stages and increased sharply to a higher level at 3- to 6-leaf stages in the bud of JF173 than that of JF914. *Ghir_D03G012390* reached the highest expression level in the buds of JF173 and JF914 at 3- and 5-true leaf stages, respectively. *Ghir_D03G012430* is a *pan1* gene. As reported, *pan1* functions in cell asymmetric division and development (Best et al., 2021; Lu et al., 2022). *Ghir_D03G012390* codes a FAM214B protein, which is vital in cell aging (Hernandez-Segura et al., 2017; Macedo et al., 2018). As JF173 has lower NFFB and better early maturity, the different expression patterns of *Ghir_D03G012430* and *Ghir_D03G012390* imply that they may be involved in NFFB formation and earliness regulation in cotton.

Data availability statement

The data presented in the study are deposited in the SRA repository, accession number PRJNA821354.

Author contributions

ML and XJ: conceived the project and set the scientific objectives. JZ, HZ, GW, and SW contributed to equipment preparation and data acquisition. XJ: wrote the manuscript. ML and GW: reviewed and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by Basic research funds of the Hebei Academy of Agriculture and Forestry Sciences (2021060206), the National Natural Science Foundation of China (32201758), Hebei Modern Agricultural Industry Technology System Innovation Team Construction Project - Mechanic Picked Variety Breeding (HBCT2018040202), and HAAFS Science and Technology Innovation Project (2022KJCXZX-LYS-14).

Acknowledgments

We thank the staff of Shanghai Majorbio Bio-pharm Technology Co., Ltd. (Shanghai, China) for their support during the genomic data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1113059/full#supplementary-material

References

Best, N. B., Addo-Quaye, C., Kim, B., Weil, C. F., Schulz, B., Johal, G., et al. (2021). Mutation of the nuclear pore complex component, aladin1, disrupts asymmetric cell division in *Zea mays* (maize). *G3 Genes|Genomes|Genetics* 11, jkab106. doi: 10.1093/g3journal/jkab106

Cheng, S., Chen, P., Su, Z., Ma, L., Hao, P., Zhang, J., et al. (2021). High-resolution temporal dynamic transcriptome landscape reveals a *GhCAL*-mediated flowering regulatory pathway in cotton (*Gossypium hirsutum* l.). *Plant Biotechnol. J.* 19, 153–166. doi: 10.1111/pbi.13449

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Chen, Z., Tang, D., Hu, K., Zhang, L., Yin, Y., Ni, J., et al. (2021). Combining QTL-seq and linkage mapping to uncover the genetic basis of single vs. paired spikelets in the advanced populations of two-ranked maize×teosinte. *BMC Plant Biol.* 21, 572. doi: 10.1186/s12870-021-03353-3

Chen, W., Yao, J., Chu, L., Yuan, Z., Li, Y., and Zhang, Y.. (2015). Genetic mapping of the nulliplex-branch gene (*gb_nb1*) in cotton using next-generation sequencing. *Theor. Appl. Genet.* 128, 539–547. doi: 10.1007/s00122-014-2452-2

Chen, D., Zhou, X., Chen, K., Chen, P., Guo, J., Liu, C., et al. (2022). Fine-mapping and candidate gene analysis of a major locus controlling leaf thickness in rice (*Oryza sativa* l.). *Mol. Breeding* 42, 6. doi: 10.1007/s11032-022-01275-y

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2014). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Gao, J., Shi, Y., Wang, W., Wang, Y., Yang, H., Shi, Q., et al. (2021). Genome sequencing identified novel mechanisms underlying virescent mutation in upland cotton *Gossypiuma hirsutum*. *BMC Genomics* 22, 1–10. doi: 10.1186/s12864-021-07810-z

Godoy, A., and Palomo, G. (1999). Genetic analysis of earliness in upland cotton (*Gossypium hirsutum* l.). i. morphological and phenological variables. *Euphytica* 105, 155–160. doi: 10.1023/A:1003490016166

Guo, Y., McCarty, J. C., Jenkins, J. N., An, C., and Saha, S. (2009). Genetic detection of node of first fruiting branch in crosses of a cultivar with two exotic accessions of upland cotton. *Euphytica* 166, 317–329. doi: 10.1007/s10681-008-9809-z

Guo, Y., McCarty, J. C., Jenkins, J. N., and Saha, S. (2008). QTLs for node of first fruiting branch in a cross of an upland cotton, *Gossypium hirsutum* l., cultivar with primitive accession Texas 701. *Euphytica* 163, 113–122. doi: 10.1007/s10681-007-9613-1

Han, Z., Hu, Y., Tian, Q., Cao, Y., Si, A., Si, Z., et al. (2020). Genomic signatures and candidate genes of lint yield and fibre quality improvement in upland cotton in xinjiang. *Plant Biotechnol. J.* 18, 2002–2014. doi: 10.1111/pbi.13356

Hernandez-Segura, A., de Jong, T. V., Melov, S., Guryev, V., Campisi, J., Demaria, M., et al. (2017). Unmasking transcriptional heterogeneity in senescent cells. *Curr. Biol.* 27, 2652–2660. doi: 10.1016/j.cub.2017.07.033

Hu, Y., Chen, J., Fang, L., Zhang, Z., Ma, W., Niu, Y., et al. (2019). *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat. Genet.* 51, 739–748. doi: 10.1038/s41588-019-0371-5

Hu, Z., Shi, X., Chen, X., Zheng, J., Zhang, A., Wang, H., et al. (2022). Fine-mapping and identification of a candidate gene controlling seed coat color in melon (*Cucumis melo* l. var. chinensis pangalo). *Theor. Appl. Genet.* 135, 803–815. doi: 10.1007/s00122-021-03999-5

Illa-Berenguer, E., Van Houten, J., Huang, Z., and van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theor. Appl. Genet.* 128, 1329–1342. doi: 10.1007/s00122-015-2509-x

Jia, X., Pang, C., Wei, H., Wang, H., Ma, Q., Yang, J., et al. (2016). High-density linkage map construction and QTL analysis for earliness-related traits in *Gossypium hirsutum* l. *BMC Genomics* 17, 909. doi: 10.1186/s12864-016-3269-y

Lacape, J., Gawrysiak, G., Cao, T., Viot, C., Llewellyn, D., Liu, S. , et al. (2013). Mapping QTLs for traits related to phenology, morphology and yield components in an inter-specific *Gossypium hirsutum*×*G. barbadense* cotton RIL population. *Field Crops Res.* 144, 256–267. doi: 10.1016/j.fcr.2013.01.001

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R., et al. (2015). Genome sequence of cultivated upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. *Nat. Biotechnol.* 33, 524–530. doi: 10.1038/nbt.3208

Li, C., Fu, Y., Liu, Q., Du, L., and Trotsenko, V. (2020). A review of genetic mechanisms of early maturity in cotton (*Gossypium hirsutum* l.). *Euphytica* 216, 120. doi: 10.1007/s10681-020-02656-0

Liu, H., Zhang, L., Mei, L., Quampah, A., He, Q., Zhang, B., et al. (2020). *qOil-3*, a major QTL identification for oil content in cottonseed across genomes and its candidate gene analysis. *Ind. Crops Products* 145, 112070. doi: 10.1016/j.indcrop.2019.112070

Li, C., Wang, Y., Ai, N., Li, Y., and Song, J. (2018). A genome-wide association study of early-maturation traits in upland cotton based on the CottonSNP80K array. *J. Integr. Plant Biol.* 60, 970–985. doi: 10.1111/jipb.12673

Li, C., Wang, C., Dong, N., Wang, X., Zhao, H., Converse, R., et al. (2012). QTL detection for node of first fruiting branch and its height in upland cotton (*Gossypium hirsutum* l.). *Euphytica* 188, 441–451. doi: 10.1007/s10681-012-0720-2

Li, C., Wang, X., Dong, N., Zhao, H., Xia, Z., Wang, R., et al. (2013). QTL analysis for early-maturing traits in cotton using two upland cotton (*Gossypium hirsutum* l.) crosses. *Breed. Sci.* 63, 154–163. doi: 10.1270/jsbbs.63.154

Li, H., Ye, G., and Wang, J. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics* 175, 361–374. doi: 10.1534/genetics.106.066811

Li, L., Zhang, C., Huang, J., Liu, Q., Wei, H., Wang, H., et al. (2021). Genomic analyses reveal the genetic basis of early maturity and identification of loci and candidate genes in upland cotton (*Gossypium hirsutum* l.). *Plant Biotechnol. J.* 19, 109–123. doi: 10.1111/pbi.13446

Li, L., Zhao, S., Su, J., Fan, S., Pang, C., Wei, H., et al. (2017). High-density genetic linkage map construction by F_2 populations and QTL analysis of early-maturity traits in upland cotton (*Gossypium hirsutum* l.). *PloS One* 12, e182918. doi: 10.1371/journal.pone.0182918

Lu, H., Lam, S., Zhang, D., Hsiao, Y., Li, B., Niu, S., et al. (2022). *R2R3-MYB* genes coordinate conical cell development and cuticular wax biosynthesis in phalaenopsis aphrodite. *Plant Physiol.* 188, 318–331. doi: 10.1093/plphys/kiab422

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., et al. (2014). QTL-seq identifies an early flowering QTL located near *Flowering locus t* in cucumber. *Theor. Appl. Genet.* 127, 1491–1499. doi: 10.1007/s00122-014-2313-z

Macedo, J. C., Vaz, S., Bakker, B., Ribeiro, R., Bakker, P. L., Escandell, J., et al. (2018). *FoxM1* repression during human aging leads to mitotic decline and aneuploidy-driven full senescence. *Nat. Commun.* 9, 2834. doi: 10.1038/s41467-018-05258-6

Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., et al. (2018). Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.* 50, 803–813. doi: 10.1038/s41588-018-0119-7

Ma, J., Pei, W., Ma, Q., Geng, Y., Liu, G., Liu, J., et al. (2019). QTL analysis and candidate gene identification for plant height in cotton based on an interspecific backcross inbred line population of *Gossypium hirsutum* × *Gossypium barbadense*. *Theor. Appl. Genet.* 132, 2663–2676. doi: 10.1007/s00122-019-03380-7

Ma, Q., Qu, Z., Wang, X., Qiao, K., Mangi, N., and Fan, S.. (2020). *EMBRYONIC FLOWER2B*, coming from a stable QTL, represses the floral transition in cotton. *Int. J. Biol. Macromolecules* 163, 1087–1096. doi: 10.1016/j.ijbiomac.2020.07.116

Ma, Z., Zhang, Y., Wu, L., Zhang, G., Sun, Z., Li, Z., et al. (2021). High-quality genome assembly and resequencing of modern cotton cultivars provide resources for crop improvement. *Nat. Genet.* 53, 1385–1391. doi: 10.1038/s41588-021-00910-2

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-

generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Paterson, A. H., Brubaker, C. L., and Wendel, J. F. (1993). A rapid method for extraction of cotton (*Gossypium* spp.) genomic DNA suitable for RFLP or PCR analysis. *Plant Mol. Biol. Reporter* 11, 122–127. doi: 10.1007/BF02670470

Ray, L., and Richmond, T. (1966). Morphological measures of earliness of crop maturity in cotton. *Crop Sci.* 6, 527–531. doi: 10.2135/cropsci1966.0011183X000600060008x

Richmond, T., and Radwan, S. (1962). A comparative study of seven methods of measuring earliness of crop maturity in cotton. *Crop Sci.* 2, 397–400. doi: 10.2135/cropsci1962.0011183X000200050010x

Su, J., Pang, C., Wei, H., Li, L., Liang, B., Wang, C., et al. (2016). Identification of favorable SNP alleles and candidate genes for traits related to early maturity *via* GWAS in upland cotton. *BMC Genomics* 17, 687. doi: 10.1186/s12864-016-2875-z

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., et al. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* 74, 174–183. doi: 10.1111/tpj.12105

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wen, T., Liu, C., Wang, T., Wang, M., Tang, F., and He, L.. (2021). Genomic mapping and identification of candidate genes encoding nulliplex-branch trait in sea-island cotton (*Gossypium barbadense* l.) by multi-omics analysis. *Mol. Breeding* 41, 1–12. doi: 10.1007/s11032-021-01229-w

Ye, S., Yan, L., Ma, X., Chen, Y., Wu, L., Ma, T., et al. (2022). Combined BSA-seq based mapping and RNA-seq profiling reveal candidate genes associated with plant architecture in *Brassica napus*. *Int. J. Mol. Sci.* 23, 2472. doi: 10.3390/ijms23052472

Zhang, T., Hu, Y., Jiang, W., Fang, L., Guan, X., Chen, J., et al. (2015). Sequencing of allotetraploid cotton (*Gossypium hirsutum* l. acc. TM-1) provides a resource for fiber improvement. *Nat. Biotechnol.* 33, 531–537. doi: 10.1038/nbt.3207

Zhang, J., Jia, X., Guo, X., Wei, H., Zhang, M., Wu, A., et al. (2021). QTL and candidate gene identification of the node of the first fruiting branch (NFFB) by QTL-seq in upland cotton (*Gossypium hirsutum* l.). *BMC Genomics* 22, 882. doi: 10.1186/s12864-021-08164-2

Zhang, B., Qi, F., Hu, G., Yang, Y., Zhang, L., Meng, J., et al. (2021). BSA-Seq-based identification of a major additive plant height QTL with an effect equivalent to that of semi-dwarf 1 in a large rice F_2 population. *Crop J.* 9, 1428–1437. doi: 10.1016/j.cj.2020.11.011

Zhang, X., Wang, G., Dong, T., Chen, B., Du, H., Li, C., et al. (2019). High-density genetic map construction and QTL mapping of first flower node in pepper (*Capsicum annuum* l.). *BMC Plant Biol.* 19, 167. doi: 10.1186/s12870-019-1753-7

Zhang, Z., Wei, T., Zhong, Y., Li, X., and Huang, J. (2016). Construction of a high-density genetic map of *Ziziphus jujuba* mill. using genotyping by sequencing technology. *Tree Genet. Genomes* 12, 76. doi: 10.1007/s11295-016-1032-9

Zhao, H., Chen, Y., Liu, J., Wang, Z., Li, F., and Ge, X.. (2022). Recent advances and future perspectives in early-maturing cotton research. *New Phytol.* doi: 10.1111/nph.18611

Zhou, Z., Zhang, C., Zhou, Y., Hao, Z., Wang, Z., Zeng, X., et al. (2016). Genetic dissection of maize plant architecture with an ultra-high density bin map based on recombinant inbred lines. *BMC Genomics* 17, 178. doi: 10.1186/s12864-016-2555-z

Zhu, J., Chen, J., Gao, F., Xu, C., Wu, H., Chen, K., et al. (2017). Rapid mapping and cloning of the *virescent-1* gene in cotton by bulked segregant analysis-next generation sequencing and virus-induced gene silencing strategies. *J. Exp. Botany* 68, 4125–4135. doi: 10.1093/jxb/erx240

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 02 February 2023
DOI 10.3389/fpls.2023.1107314

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Jinping Zhao,
Texas A&M University, United States

*CORRESPONDENCE
Liping Xu
xlpmail@126.com
Youxiong Que
queyouxiong@126.com

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 24 November 2022
ACCEPTED 23 January 2023
PUBLISHED 02 February 2023

CITATION
Lu G, Wang Z, Pan Y-B, Wu Q, Cheng W, Xu F, Dai S, Li B, Que Y and Xu L (2023) Identification of QTLs and critical genes related to sugarcane mosaic disease resistance.
Front. Plant Sci. 14:1107314.
doi: 10.3389/fpls.2023.1107314

Identification of QTLs and critical genes related to sugarcane mosaic disease resistance

Guilong Lu[1,2], Zhoutao Wang[1], Yong-Bao Pan[3], Qibin Wu[1], Wei Cheng[1], Fu Xu[1], Shunbin Dai[1], Boyu Li[1], Youxiong Que[1]* and Liping Xu[1]*

[1]Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture and Rural Affairs, Fujian Agriculture and Forestry University, Fuzhou, China, [2]Institute of Vegetables, Tibet Academy of Agricultural and Animal Husbandry Sciences, Lhasa, China, [3]USDA-ARS, Sugarcane Research Unit, Houma, LA, United States

Mosaic viral diseases affect sugarcane productivity worldwide. Mining disease resistance-associated molecular markers or genes is a key component of disease resistance breeding programs. In the present study, 285 F_1 progeny were produced from a cross between Yuetang 93-159, a moderately resistant variety, and ROC22, a highly susceptible variety. The mosaic disease symptoms of these progenies, with ROC22 as the control, were surveyed by natural infection under 11 different environmental conditions in the field and by artificial infections with a mixed *sugarcane mosaic virus* (SCMV) and *sorghum mosaic virus* (SrMV) inoculum. Analysis of consolidated survey data enabled the identification of 29 immune, 55 highly resistant, 70 moderately resistant, 62 susceptible, and 40 highly susceptible progenies. The disease response data and a high-quality SNP genetic map were used in quantitative trait locus (QTL) mapping. The results showed that the correlation coefficients (0.26~0.91) between mosaic disease resistance and test environments were significant ($p < 0.001$), and that mosaic disease resistance was a highly heritable quantitative trait ($H^2 = 0.85$). Seven mosaic resistance QTLs were located to the SNP genetic map, each QTL accounted for 3.57% ~ 17.10% of the phenotypic variation explained (PVE). Furthermore, 110 pathogen response genes and 69 transcription factors were identified in the QTLs interval. The expression levels of nine genes (*Soffic.07G0015370-1P*, *Soffic.09G0015410-2T*, *Soffic.09G0016460-1T*, *Soffic.09G0016460-1P*, *Soffic.09G0017080-3C*, *Soffic.09G0018730-3P*, *Soffic.09G0018730-3C*, *Soffic.09G0019920-3C* and *Soffic.03G0019710-2C*) were significantly different between resistant and susceptible progenies, indicating their key roles in sugarcane resistance to SCMV and SrMV infection. The seven QTLs and nine genes can provide a certain scientific reference to help sugarcane breeders develop varieties resistant to mosaic diseases.

KEYWORDS

sugarcane (*Saccharum* spp. hybrids), sugarcane mosaic disease, QTL mapping, gene mining, expression profiles

Introduction

Sugarcane mosaic disease (SMD) is a worldwide issue that has long plagued sugarcane production. The disease is mainly caused by single or co-infection of *Sugarcane mosaic virus* (SCMV), *Sorghum mosaic virus* (SrMV), and *Sugarcane streak mosaic virus* (SCSMV) (Lu et al., 2021). SMD exhibiting typical "mosaic" symptoms (Grisham, 2011) can seriously reduce the photosynthetic capacity (Bagyalakshmi et al., 2019), yield, and quality of sugarcane (Singh et al., 2003; Viswanathan and Balamuralikrishnan, 2005). Pandemic SMD has occurred many times in history and caused huge economic losses and even bankruptcies to many sugar companies (Koike and Gillaspie, 1989; Grisham, 2011). Breeding and rationally planting of SMD-resistant varieties are the most economical and effective methods to prevent and control the disease.

So far, both natural infection disease surveys and artificial inoculation-induced infection disease surveys are used in SMD resistance assessments. Using the natural infection method, Li et al. (2013); Da-Silva et al. (2015a); Yang et al. (2020), and Lavín-Castaeda et al. (2020) successively screened sugarcane breeding materials, cultivars, or hybrid offspring populations. A few varieties (lines) with immunity or good resistance to SMD provided good material for mosaic disease resistance gene mining and hybrid breeding. Although this method is simple and saves labor and time, it requires a high level of professional ability and is often affected by environments. Alternatively, several artificial inoculation methods, including friction (Da-Silva et al., 2015b; De-Souza et al., 2017), spray (Dean, 1960), stalk cutting (Li et al., 2013; Li et al., 2018), and injection inoculations (Zhou, 2015), can be well controlled and be evaluated under a set stress. Roossinck (2015) assumed that the occurrence and prevalence of plant diseases depended on a compound effect among host plants, pathogens, and environmental factors. Therefore, it is of vital importance to choose the most suitable growth stage and the optimum inoculation methods for improved accuracy of resistant phenotype identification during field evaluation.

The development of practical molecular markers and related detection methodology are the basis for molecular marker-assisted breeding. Currently, traditional DNA markers, such as amplified fragment length polymorphism (AFLP), restriction fragment length polymorphism (RFLP), and simple sequence repeats (SSR), are being used in quantitative trait locus (QTL) mapping or bulk segregant analysis (BSA) research (Xia et al., 1999; Duble et al., 2000; Xu et al., 2000; Dussle et al., 2003; Yuan et al., 2004). Several SCMV-resistance markers were identified in corn (*Zea mays* L., $2n = 2x = 20$; genome size ~2,300 Mb) (Schnable et al., 2009). Single nucleotide polymorphisms (SNP) markers are superior markers due to wide distribution, huge quantity, high stability, strong representativeness, and bi-allelicity (Rafalski, 2002). SNP chips represent a high-throughput, automated, and relatively cost-effective genotyping method (Laframboise, 2009), which has been used to identify resistance genes to *Bean common mosaic virus* in soybean ($2n = 2x = 40$) (Bello et al., 2014) and to *Soil-borne wheat mosaic virus* in wheat ($2n = 6x = 42$) (Liu et al., 2014). However, due to the complexity of the sugarcane genome ($2n = 12x = 100$~130 and genome size ~10 Gb) (Roach, 1989; D'Hont et al., 1998), sequencing technology, and high cost, only two SNP chips, namely, the 345K chip of Aitken et al. (2017) and the 100K chip of You et al. (2019), have been developed in sugarcane. The 100K SNP chip has a polymorphism rate of up to 77.04% and has been successfully used in QTL mapping of disease resistance markers to yellow leaf disease (You et al., 2019), ratoon stunting disease (You et al., 2020), and leaf blight disease (Wang et al., 2021) in sugarcane.

In plants, compared to qualitative resistance traits, quantitative resistance traits are more broad-spectrum and persistent and play an important role in preventing large-scale disease outbreaks due to the loss of a single gene resistance (Poland et al., 2009). For instance, a QTL locus *qMdr9.02* was found to be associated with resistance to southern leaf blight, northern leaf blight, and gray leaf spot in maize (Yang et al., 2017). However, to date, only four SCMV resistance-associated markers (*AFLP-346*, *AFLP-372*, *AFLP-538*, and *CV29.13*), each accounting for 5.51 to 14.02% of PVE, were reported by Burbano et al. (2022). The objectives of this study were to construct a genetic mapping population, to evaluate the SMD response of the mapping population, and to develop SMD resistance-associated QTL markers and suggest candidate genes for the improvement of the efficiency and accuracy of sugarcane breeding.

Materials and methods

Plant material and field planting

Two hundred and eighty-five F_1 progeny were produced from a cross between YT93-159 (moderately resistant to SMD) and ROC22 (highly susceptible to SMD). The cross was made in 2014 at the Hainan Sugarcane Breeding Station, Yacheng, Hainan, China. After vegetative propagation, stems of these progeny were planted at five different ecological sites, namely, Cangshan (119°14'E, 26°5'N), Longchuan (97°53'E, 24°15'N), Suixi (110°10'E, 21°6'N), Tianyang (107°0'E, 23°39'N), and Yuanjiang (101°59'E, 23°36'N) (Figure 1; Supplementary Table 1). A randomized block design was adopted for field planting. Specifically, the trial design in Cangshan and Longchuan contained three replications, Suixi and Yuanjiang contained two replications, and Tianyang contained one replicate. Specific row spacing and planting density were shown in Supplementary Table 2. The five ecological sites were routinely managed according to conventional planting measures, and stalk-cutting was done at the end of December each year.

Mosaic disease survey

By natural infection

To identify the appropriate survey season, SMD symptoms on a field grown, highly susceptible progeny FN14-255 were monitored monthly on the campus of Fujian Agriculture and Forestry University (FAFU) (119°14'E, 26°5'N). Three typical +1 leaves were sampled for comparison. The three periods showing the most severe symptoms were selected for investigating natural SMD incidence.

By artificial inoculation

Before planting, a machete was used to cut the stem of FN14-255 into single-bud pieces, which were rinsed in running water overnight. Only single-bud pieces that met the criteria of 1) having one full and

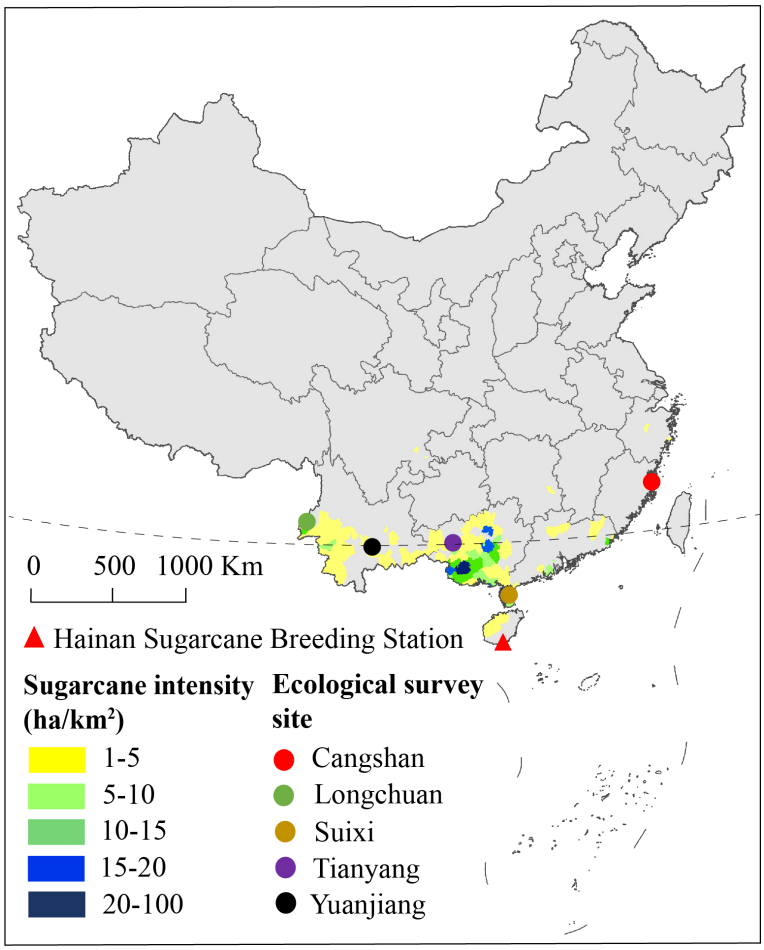


FIGURE 1
Ecological survey sites and sugarcane crop density in China based on the data from the 2020 Statistical Yearbook.

healthy bud, and 2) with flat incisions without any cracks were kept. A super constant temperature tank (Ningbo Prandt Instrument Co., Ltd, Ningbo, China) was used for hot water treatment. Water temperature was set and kept at ± 0.2°C of 50°C (CK), 55°C, 57°C, 59°C, and 61°C. Water level was maintained at about 2/3 tank full. Treatment was for 30 minutes. Once the treatment was completed, the stems were rinsed in running water until the buds cooled completely. The buds were cultured in a greenhouse (Supplementary Figure 1) under 12 h light/12 h dark with a light intensity of 15,000 Lx and a relative humidity of 60%. Greenhouse temperature was set to 28°C before inoculation and 25°C after inoculation. Each treatment had 30 buds with three replications. After 30 d, the one-step multiplex reverse transcription PCR (RT-PCR) method of Shan et al. (2020) was used to detect different sugarcane mosaic virus. The oligonucleotide sequence of species-specific RT-PCR primers and the length of targeted fragments are shown in Table 1.

The method of Li et al. (2013) was used to configure the viral inoculum mixture. The viral source was SMD symptomatic leaves from sugarcane variety Funong 41 that was planted on the Sugarcane Farm on the campus of FAFU. SCMV and SrMV pathogens were detected in these leaves by RT-PCR (Supplementary Figure 2). YT93-159 and ROC22 were used to test different inoculation methods, including spray, micro-injection, quartz sand friction, abrasive cloth friction, rasp friction, young stem cut, single bud soaking, single bud soaking and quartz sand friction (Supplementary Figure 3, Supplementary Table 3), and to choose the best inoculation method to inoculate the test population. In 2021, three batches of viral inoculums were administered successively. One was conducted at the sugarcane station of FAFU during February to April. Another was conducted in a climate-controlled greenhouse of the Key Laboratory of Sugarcane Biology and Genetic Breeding, Ministry of Agriculture and Rural Affairs, FAFU from May to July. A final inoculation was conducted in the same greenhouse from October to December. For each genotype, 15 single buds were inoculated with three replications and were kept in the dark for 24 h after inoculation. Four weeks post inoculation, SMD incidence was investigated for three consecutive sessions with an interval of one week.

Resistance evaluation

One growth cycle at one ecological site and a batch of artificial inoculation treatments were considered as one environment. The

TABLE 1 Species-specific RT-PCR primers for the detection of three sugarcane mosaic viruses.

Virus	Primer sequence (5′→3′)	Annealing temperature (°C)	Amplification size (bp)
SCMV	F: GCGCGGTATGCATTTGACTT	58	200
	R: CACTCCCAACAGAGAGTGCAT		
SrMV	F: AACAGGATGCCGATGCGAAA		450
	R: CGTTGATGTTCGGTGAGCAA		
SCSMV	F: GAACGCAGCCACCTCAGAAT		800
	R: CCAAAATGAGCGCCTCCGAT		

highest SMD incidence rate out of the three surveys was used to determine the level of SMD resistance for each F_1 progeny in a single environment. Comprehensive evaluation was based on the maximum value of resistance across multiple natural and artificial inoculation infection environments. The SMD grading system was set according to the method of Li et al. (2000) (Table 2). During comprehensive evaluation, if the disease incidence rate of ROC22 (control) in an environment was more than 66.01%, the external SMD stress was considered sufficient, and the survey data valid. If the disease incidence rate of ROC22 (control) in an environment was less than 66.01%, then the external SMD stress was assumed to be insufficient, and the environmental data discarded. The following formula was used to calculate SMD incidence rate (%):

SMD incidence rate (%) = number of diseased plants/total number of plants per F_1 progeny × 100%.

Correlation analysis and generalized heritability estimation

The QTL IciMapping V4.2 software (Chinese Academy of Agricultural Sciences, Beijing, China) was used to analyze the correlation and calculate the generalized heritability (H^2) using the following calculation formula:

$$H^2 = \sigma_g^2/(\sigma_g^2 + \frac{\sigma_{ge}^2}{n} + \frac{\sigma_e^2}{nr}),$$

Where σ_g^2 is genotype variance, σ_e^2 is error variance, σ_{ge}^2 is genotype-by-environment interaction variance, n is the number of environments; and r is number of survey periods within each environment.

TABLE 2 Resistance grading based on SMD incidence.

Grade	Resistance	SMD Incidence (%)
1	Immune	0
2	Highly resistant	0.01~10.00
3	Moderately resistant	10.01~33.00
4	Susceptible	33.01~66.00
5	Highly susceptible	66.01~100

QTL mapping

The SMD resistance grading data of the F_1 progeny population and the sugarcane 100K SNP chip-based genetic map (Supplementary Table 4) (Wang et al., 2021) were used to conduct QTL mapping using the inclusive composite interval mapping (ICIM) of GACD 1.2 software (Chinese Academy of Agricultural Sciences, Beijing, China), with a logarithm of odds (LOD) threshold of 2.5 and other default parameters. Loci with ≥ 10% phenotypic variation explained (PVE) values were defined as major QTLs, and loci with< 10% PVE were minor QTLs. QTLs were named according to McCouch et al. (1997) with "*q*" plus the sugarcane mosaic disease resistance (Rsm) trait, followed by linkage group number in italics. R software (R-Tools Technology Inc., Ontario, Canada), Origin 9.0 software (OriginLab Inc., Massachusetts, USA), and Adobe Illustrator CS6 software (Adobe Systems Inc., California, USA) were used to draw the position of QTL on the linkage group.

Candidate gene mining

The protein sequences of all genes in the QTL interval were extracted according to the GFF annotation file of a *Saccharum officinarum* genome (http://sugarcane.zhangjisenlab.cn/sgd/html/-index.html). The Plant Pathogen Receptor Genes database (PRGdb 4.0, http://prgdb.org/prgdb4/) was used to search for genes related to disease resistance. At the same time, disease resistance-related transcription factors were extracted from the plant transcription factor database (TFDB 5.0, http://planttfdb.gao-lab.org/index.php) (Osuna-Cruz et al., 2018).

Critical gene and functional structure prediction

Stems of Yuetang 93-159, ROC22, five immune, and five highly susceptible progeny were detoxified in a hot water bath as previously described. Plants with 2~3 fully expanded leaves from the detoxified buds were inoculated with a mixed inoculum of SCMV and SrMV by quartz sand friction. Leaf samples were taken on 0 d, 1 d, and 4 d post inoculation, RT-PCR was conducted to detect the viruses at 4 d post inoculation (Supplementary Figure 4). There were four plants in each of the three biological replicates. RNA was extracted by the Trizol method, and the integrity of the extracted RNA samples was checked

using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The integrity number of a qualified RNA sample was considered greater than 6.0, and the detection quality was A-level (Supplementary Table 5). The qualified RNA samples were sent to Novogene Bioinformatics Technology Co., Ltd. (Beijing, China) for transcriptome sequencing. The DNBSEQ-T7 (Shenzhen Huada Intelligent Technology Co., Ltd., Shenzhen, China) sequencing platform was used for paired-end sequencing, and each library yielded ≥ 12 Gb of sequence data (Supplementary Table 6). The Transcripts Per Kilobase Million (TPM) normalization method (Wang et al., 2021) was used to calculate the expression levels of all genes. TBtools V1.0986 software (South China Agricultural University, Guangzhou, China) was used to draw an expression heat map of candidate genes, and to locate significantly differentiated key genes in the *S. officinarum* genome. An online tool GSDS V2.0 (http://gsds.gao-lab.org/) was used to describe the gene structure. The *Arabidopsis* genome (https://www.arabidopsis.org/Blast/index.jsp) was referred for functional annotation with e-value threshold set to $1e^{-10}$.

Data statistics and analysis

A Canon EOS 60D camera (Canon Inc., Tokyo, Japan) was used to capture the images of SMD symptoms. Data were achieved as Excel 2010 (Microsoft Inc., Washington, USA) spreadsheets. Duncan's significant difference test and descriptive statistics were performed using IBM SPSS® V25 software (International Business Machines Inc., California, USA).

Results

Phenotypic analysis and evaluation

Determination of the natural survey period

The SMD symptoms of a highly susceptible progeny (FN14-255) are shown in Figure 2. The figure shows the symptoms of infected sugarcane leaves were more clearly distinguishable during February to April and October to December, with mosaic symptoms covering the entire leaf. Nevertheless, the symptoms were significantly weakened in January and in May to September, especially from June to August, the symptoms were suppressed by high temperature, and can only be observed at the bottom of the leaves. Therefore, the field natural incidence survey was arranged in March, April and November, respectively.

Hot-water detoxification and artificial inoculation

Germination time was obviously delayed, and germination rate was significantly reduced with increasing hot water temperature (Supplementary Table 7). On the other hand, mild leaf symptoms could be seen from the 50°C treatment. And even barely visible from the 55°C treatment. However, no symptom was observable from the 57°C, 59°C, and 61°C treatments. As shown in Supplementary Figure 5, no band was visible on the gels, indicating that all three target viruses were not detectable for the samples treated at 59°C and 61°C. Therefore, a hot water treatment at 59°C for 30 min can completely detoxify the viruses, albeit with a germination rate of about 30% (Supplementary Table 7). The 'single bud soaking + quartz sand friction' method had the highest inoculation efficiency (Supplementary Table 8). Therefore, this method was used to inoculate the mapping population material.

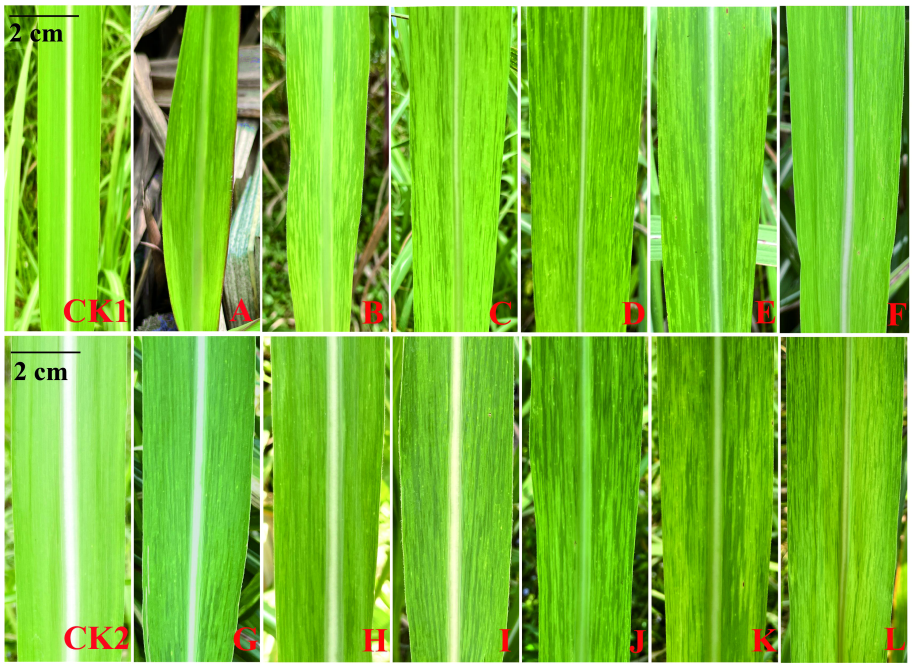


FIGURE 2
SMD symptoms of a highly susceptible progeny (FN14-255) observed in different months CK1: disease free control (January); CK2: disease free control (December); **(A–L)**: January-December, respectively.

Comprehensive evaluation

The SMD survey data for the F_1 mapping population from 11 natural infection and 3 artificial inoculation infection environments during 2020 to 2022 are shown in Supplementary Table 9. The frequency distribution of SMD resistance grades within this population in 14 environments is shown in Supplementary Figure 6. The data from two environments at Cangshan ecological site (block 2) were excluded from the comprehensive evaluation and resistance analysis due to insufficient pathogen stress. Accordingly, the population was comprehensively evaluated based on nine natural environments and three artificial inoculation environments. Among the 285 progenies, 29 immune, 55 highly resistant, 70 resistant, 62 susceptible, and 40 highly susceptible progenies were identified. The remaining 29 progenies had inconsistent SMD responses. The SMD resistance trait segregated widely within the F_1 mapping population and showed an obvious hybrid vigor (Heterosis) phenomenon (Figure 3). That was in line with the typical characteristics of a quantitative trait, indicating its suitability for QTL analysis.

Correlation analysis and generalized heritability

Certain differences of SMD incidence were observed in the mapping population across different environments. For example, SMD incidence in the ratoon crop was significantly higher than the plant cane crop. The SMD tended to accumulate when the sugarcane crop underwent prolonged ratooning. Correlation coefficients between the resistance trait and different environments were 0.26~0.91 (Supplementary Table 10), all these values were very significant ($p< 0.001$), indicating that the SMD resistance was a stable trait. Not surprisingly, the estimated broad sense heritability (H^2) of SMD resistance in this mapping population under 14 environments was 0.85, which implied that the SMD resistance trait was mainly determined by genetic factors.

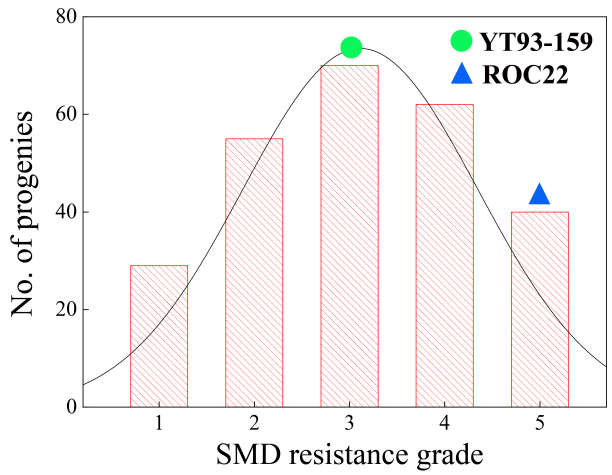


FIGURE 3
Distribution of five SMD resistance grades (Table 2) within a sugarcane mapping population YT93-159 belongs to Grade 3 and ROC22 belongs to Grade 5.

QTL mapping

Seven SMD resistance-related QTLs were detected (Table 3), which could explain 46.53% of the PVE. One major QTL, *qRsm-Y12*, could explain 17.10% of the PVE. The other six were minor QTLs, each could explain 3.57% ~ 7.70% of PVE. Four QTLs were detected on the YT93-159 map, and the remaining three QTLs were detected on the ROC22 map (Figure 4). The maximum genetic distance of each QTL from the nearest marker was 2.4 cM, the minimum was 0, and the average genetic distance was about 1.1 cM.

Candidate gene mining

According to the sequence information of the markers on either side of the QTL (Supplementary Table 11), 1,525 candidate genes were searched in the seven QTLs regions. In total, 110 disease resistance candidate genes were identified, whose gene products included CC-TM (coiled-coil plus transmembrane receptor), LRR (leucine rich repeats), RLK (receptor-like protein kinases), WAK (wall-associated receptor kinase), and others domain. In addition, 69 transcription factors were identified, including AP2 (APETALA2), bHLH (basic helix-loop-helix), bZIP (basic region/leucine zipper), ERF (ethylene response factor), MYB (myeloblastosis), SBP (squamosa promoter binding protein) and other types of transcription factors (Supplementary Table 12). These genes and transcription factors may directly or indirectly involve in regulating sugarcane response to mosaic virus infection.

Critical gene prediction

The gene expression levels of 110 pathogen-responsive genes and 69 transcription factors obtained by map mapping were presented in Figure 5. Among the candidate genes related to disease resistance, it was found that genes such as *Soffic.07G0015370-1P*, *Soffic.09G0016460-1T*, and *Soffic.09G0018730-3P* had significant expression differences between resistant and susceptible progenies, including three transcription factors and six pathogen response genes. These nine genes contained conserved domains such as bHLH_AtILR3_like, LRR, STKc_SNT7_plant and that were closely related to plant disease resistance (Table 4). The genomic positions, conserved domains and gene structures of the nine predicted genes are shown in Figure 6. It is speculated that these genes may be key to the resistance of sugarcane to SCMV and SrMV, and can be a focus for future research.

Discussion

Mosaic disease is one of the most important viral diseases in sugarcane and has threatened the security and sustainability of the world sugarcane industry for a long time (Wu et al., 2012). In recent years, with the increasing pressure of natural stress, the differentiation of plant viruses has accelerated (Roossinck, 2015). The genetic basis of modern sugarcane cultivars is narrow, and the utilization of resistant genes and genotypes is limited. There is an increasing chance of a large-

10.3389/fpls.2023.1107314

TABLE 3 SMD resistance-related QTLs in a F_1 progeny mapping population from the YT93-159 × ROC22 cross.

QTL	Position	Left/Right markers	LOD	PVE (%)	Effect female	Effect male	Effect FM	GD (cM)	Marker[a]	Distance (cM)[b]
qRsm-Y12	16	AX-171367442/AX-171312668	10.19	17.10	-0.01	-0.05	0.50	9.5	AX-171312668	0.9
qRsm-Y41	35	AX-171308038/AX-171265900	2.72	3.57	0.06	-0.21	-0.11	6.8	AX-171308038	1.5
qRsm-Y52	4	AX-171266761/AX-117172243	3.25	4.90	0.27	-0.04	-0.02	25.3	AX-171266761	0.4
qRsm-Y57	60	AX-171332119/AX-171288089	3.37	5.12	0.19	-0.03	0.22	5.6	AX-171288089	2.4
qRsm-R14	0	AX-171290689/AX-171329853	2.52	3.88	-0.11	-0.10	0.19	1.8	AX-171290689	0
qRsm-R23	17	AX-171330585/AX-171286409	3.44	7.70	0.11	0.13	-0.25	0.7	AX-171286409	0.2
qRsm-R92	3	AX-171360287/AX-171296656	2.62	4.26	-0.22	0.12	-0.09	5.3	AX-171296656	2.3

"Y", YT93-159; "R", ROC22; LOD, logarithm of odds; PVE, phenotypic variation explained; GD, genetic distance between left and right markers; [a] Nearest marker from the QTL peak, [b] Distance of nearest marker from the respective QTL peak.

scale epidemic of mosaic diseases. Since different sugarcane varieties may have different resistances to the virus, breeding and careful distribution of disease-resistant varieties is the most economical and effective method to control mosaic disease. Therefore, it is imperative to fully explore the specifics of germplasm resistance and expand research on resistance-related molecular markers or key genes to further improve breeding efficiency.

In this study, SMD surveys were based on the "mosaic" symptom manifested under multiple environments. The results of resistance to mosaic disease in the experimental population showed that the overall disease incidence upon artificial inoculation was significantly higher than that upon natural infection. Due to many years of sugarcane production and greater levels of pathogen pressure, the overall disease incidence in sugarcane production areas of Guangxi and Yunnan is significantly higher than other ecological regions in China (Supplementary Table 9). In our study, inconsistent SMD incidences were observed across different habitats. The pathogen pressure of SMD was not high enough on the newly planted sugarcane crop at Cangshan ecological site (block 2) in 2020 and 2021, therefore, the survey data from these two environments were discarded. Therefore, the evaluation was only carried out with the progeny with the highest level of resistance across nine natural infection environments and three artificial inoculation infection environments. Excluding 29 F_1 progeny with inconsistent levels of SMD resistance across different environments, 256 progeny of the F_1 mapping population were included in further analysis. The 29 F_1 progenies that were immune to SMD will be valuable in molecular breeding to develop SMD resistant sugarcane cultivars.

Sugarcane is a vegetatively propagated crop, and multiple sets of a mapping population can be propagated genetic research (Asnaghi et al., 2004). This study showed that the correlation coefficients among SMD resistance data sets from the various environments were highly significant ($p< 0.001$) at 0.26 ~ 0.91 (Supplementary Table 10). This indicates that SMD resistance is stable under different environmental conditions. The consolidated survey results showed that the frequency of the five grades followed a continuous normal distribution and that the Grades 1 and 2 contained 84 super-parent segregants with a better resistance level than the parent YT93-159, which is resistant to SMD (Grade 3) (Figure 3). This is in line with the typical characteristics of a quantitative trait controlled by polygenes. The generalized heritability (H^2) of the SMD resistance across different environments was 0.85, which is obviously higher than the H^2 values reported on sugar content (0.57), plant height (0.57), effective stem number (0.65), single stem weight (0.56), and yield (0.49) (Barreto et al., 2019). This may be due to the long-term accumulation and habitation of the virus in sugarcane and the less effective management of SMD than on plant yield-related traits. The SMD resistance trait is mainly controlled by genetic factors, which can be identified using the map mapping method.

Mapping population size and molecular marker density directly affect the accuracy and resolution of marker localization for the target trait (Beavis, 1994). So far, most of the sugarcane populations for QTL mapping of agronomic traits are made up of between 100 and 200 individuals with traditional markers, such as AFLP, RFLP or SSR (Raboin et al., 2006; Yang, 2015; Singh et al., 2016). Due to the lack of detection tools, high-density genotyping of large populations, the genetic distance between the QTL markers and the gene of interest is relatively large (Daugrois et al., 1996; Raboin et al., 2006). In this study, linkage analysis was performed using a high-density map constructed by the Axiom Sugarcane 100K SNP chip, which contains 100,097 low-dose SNPs with a broad genetic basis and mainly distributed in gene regions. This chip includes 64,726 single-dose markers and 35,371 double-dose markers (You et al., 2019). Furthermore, the F_1 progeny mapping population used in this study consisted of 256 eligible F_1 progeny, which is significantly more

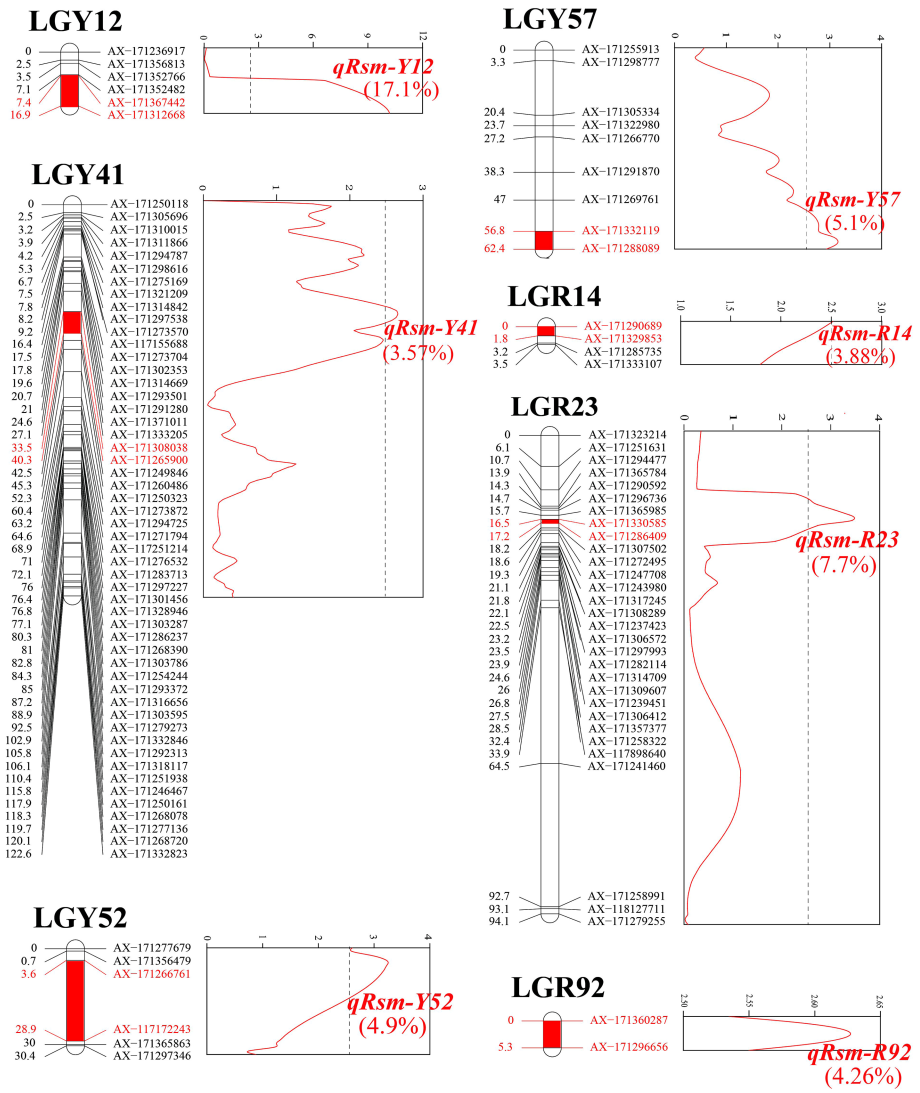


FIGURE 4
Location of seven SMD resistance-related QTLs (*q*) on sugarcane genetic linkage maps "Rsm", resistance trait to sugarcane mosaic disease; "Y", YT93-159; and "R", ROC22. The colored text values, phenotypic variation explained (PVE).

than those of previous studies (Raboin et al., 2006; Yang et al., 2015; Singh et al., 2016).

The genetic analysis of SMD resistance was analyzed in this study. Seven SMD resistance-related QTLs were detected, only one of which, *qRsm-Y12*, was a major QTL that could explain 17.1% of the PVE. The genetic effect of *qRsm-Y12* is similar to the PVE effects seen for SCMV resistance (14.02%) by marker *AFLP-346* in sugarcane (Burbano et al., 2022) and the 15.3% ~ 15.8% PVE effect of a major QTL *R-scm3* related to SCMV resistance in maize (Zhang et al., 2003). The seven QTL markers identified in this study range in distance from the nearest marker from 0 to 2.4 cM, with an average of 1.1 cM, which is similar to those seen for sugarcane brown rust resistance-associated markers (0.1 cM ~ 8.1 cM) (Yang et al., 2017) and sugarcane orange rust markers (0.2 cM ~ 2.2 cM) (Yang et al., 2018). This further demonstrated the feasibility and reliability of using SNP genetic maps to locate target trait-related QTLs. However, even with a high-quality sugarcane SNP map, the distance of the closest markers on either side of the QTL is relatively large (Wang et al., 2021). For example, the distance between QTL *qRsm-Y57* and the closest marker is 2.4 cM, which makes target trait localization difficult and highlights the need for fine localization of SNP markers.

The major disease resistance traits in plants may generally be described by a gene-for-gene mechanism. The Avr products of pathogen-encoded avirulence genes are specifically recognized directly or indirectly by specific proteins encoded by the cognate plant disease resistance genes (Flor, 1971; Jia et al., 2000; Yakupjan et al., 2015). When plants sense a pathogen invasion signal, the disease resistance genes are activated through a series of signal transmissions. During this process, transcription factors play an important role in the defensive responses. For example, they may inhibit or activate the transcriptional expression of target genes by binding to specific DNA sequences in target gene promoters (Zhang

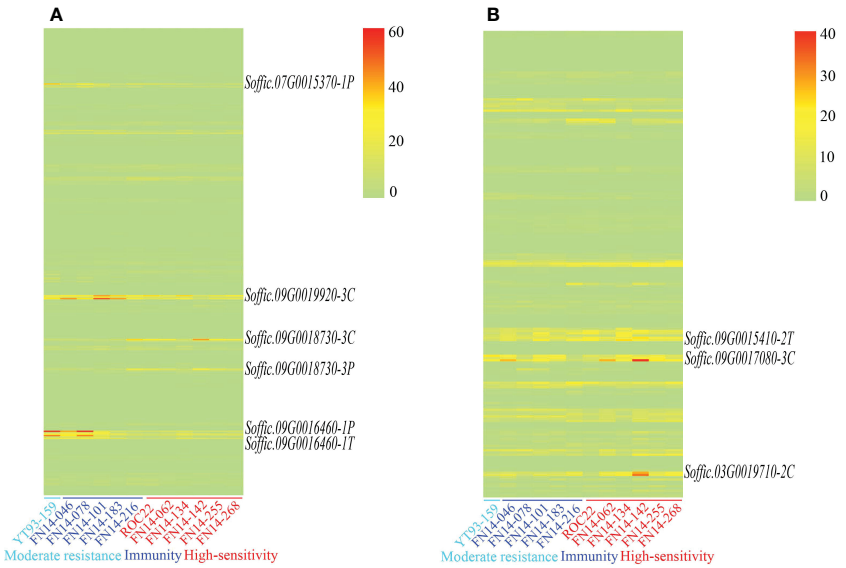


FIGURE 5
Expression of SMD resistance-related candidate genes **(A)** disease response-related genes; **(B)**. disease resistance-related transcription factors.

et al., 2009). Plant leucine-rich repeat (LRR) receptor-like protein kinases represent a large group of protein families that play important roles in disease resistance (Smakowska-Luzan et al., 2018). Zhang et al. (2021) showed that a homologous *OsRLP1* gene regulated rice resistance to *Rice black-streaked dwarf virus* infection. Qi et al. (2014) found that a mutant of *A. thaliana* protein kinase AVRPPHB susceptible (PBS1) was defective in sensing the avirulence gene *avrPphB* of *Pseudomonas syringae*. Lee et al. (2015) showed that a serine/threonine kinase domain protein encoded by *OsPBL1* gene might play a role in rice stripe resistance. Chang et al. (2022) found a FKBP-type peptidyl-prolyl *cis-trans* isomerase (PPIase) could interact with the motor protein of *Tomato leaf curl New Delhi virus*, and its transient overexpression reduced the virus replication. Aparicio and Pallás (2017) confirmed that bHLH transcription factor can promote salicylic acid-dependent defense signaling by interacting with the *Alfalfa mosaic virus* CP protein. Studies have also shown that *MdMYB73* can improve apple's resistance level to *Botryosphaeria dothidea* through the salicylic acid pathway (Gu et al., 2020). The

TABLE 4 Information for SMD resistance-related key genes.

No.	QTL	Candidate gene	*Arabidopsis* homologous gene	Conserved domain	Gene description
1	*qRsm-R14*	*Soffic.07G0015370-1P*	*AT2G43560*	FkpA super family	FKBP-like peptidyl-prolyl *cis-trans* isomerase family protein
2	*qRsm-Y52*	*Soffic.09G0015410-2T*	*AT5G54680*	bHLH_AtILR3_like	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
3	*qRsm-Y52*	*Soffic.09G0016460-1T*	*AT5G01920*	STKc_SNT7_plant	Protein kinase superfamily protein
4	*qRsm-Y52*	*Soffic.09G0016460-1P*	*AT5G01920*	STKc_SNT7_plant	Protein kinase superfamily protein
5	*qRsm-Y52*	*Soffic.09G0017080-3C*	*AT3G12480*	BUR6 super family	nuclear factor Y, subunit C11
6	*qRsm-Y52*	*Soffic.09G0018730-3P*	*AT5G25930*	LRR	kinase family with leucine-rich repeat domain-containing protein
7	*qRsm-Y52*	*Soffic.09G0018730-3C*	*AT5G25930*	LRR	kinase family with leucine-rich repeat domain-containing protein
8	*qRsm-Y52*	*Soffic.09G0019920-3C*	*AT1G68830*	PLN03225	Serine/Threonine kinase domain protein
9	*qRsm-Y52*	*Soffic.03G0019710-2C*	*AT5G23000*	PLN03091 super family	myb domain protein 37

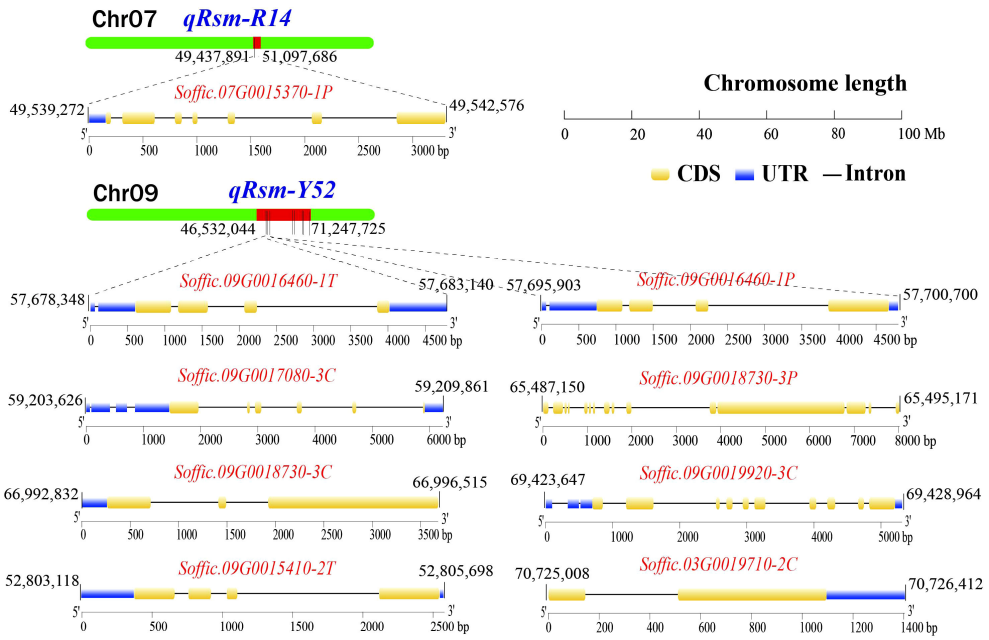


FIGURE 6
The genomic location, conserved domain, and gene structure of SMD resistance-related candidate key genes (UTR, untranslated region; CDS, coding sequence).

expression of a MYB transcription factor *CaPHL8* was upregulated in *Ralstonia solanacerum* infected pepper plants. The upregulated expression activated the expressions of immune-related genes to enhance the defense response of pepper (Noman et al., 2019). O'Conner et al. (2021) showed that overexpression of *GmNF-YC4-2* in soybean increased seed protein content, exhibited a broad disease resistance, and accelerated soybean maturation.

In this study, a total of 110 pathogen-responsive genes and 69 transcription factors were identified in the interval regions of the QTLs. Among them, nine candidate genes were obtained in the interval region of the major QTL *qRsm-Y12*, including one transcription factor and eight resistance genes. Basically, plants share a common resistance mechanism to the same type of pathogen (Jones and Dangl, 2006; Li et al., 2020). SCMV and SrMV are the most widely distributed sugarcane mosaic virus in the world, with SCSMV mainly distributed in Asia (Lu et al., 2021). Therefore, we used an artificial inoculum that only contained SCMV and SrMV. Combined with the TPM normalization results of RNA-seq gene expression after inoculation of SCMV and SrMV, six genes and three transcription factors had significantly different levels of expression between resistant and susceptible materials. Two genes, *Soffic.09G0018730-3P* and *Soffic.09G0018730-3C*, contained LRR domains. Two genes, *Soffic.09G0016460-1T* and *Soffic.09G0016460-1P*, encoded kinase superfamily proteins. Gene *Soffic.09G0019920-3C* encoded a serine/threonine kinase domain protein. Gene *Soffic.07G0015370-1P* encoded a PPIase family protein. Among the transcription factors, *Soffic.09G0015410-2T* is a bHLH transcription factor, *Soffic.03G0019710-2C* encodes a MYB transcription factor, and *Soffic.09G0017080-3C* encodes a NF-YC transcription factor. It is thus speculated that these six genes and three transcription factors may have potential functions in sugarcane mosaic disease resistance.

Conclusions

This study showed that the SMD resistance trait of 256 F_1 progeny of a cross (YT93-159 × ROC22) tested under different environments was significantly correlated ($p< 0.001$) with correlation coefficients of 0.26~0.91, and hence was a highly heritable quantitative trait ($H^2 = 0.85$). Based on the consolidated multiple data sets of SMD resistance, 29 immune, 55 highly resistant, 70 moderately resistant, 62 susceptible, and 40 highly susceptible F_1 progeny were identified. Using a high-quality SNP chip, seven SMD resistance-related QTLs were located. One major QTL, *qRsm-Y12*, explained 17.10% of the PVE and six minor QTLs, namely, *qRsm-Y41*, *qRsm-Y52*, *qRsm-Y57*, *qRsm-R14*, *qRsm-R23*, and *qRsm-R92*, explained 3.57% ~ 7.70% of the PVE. A total of 110 SMD response genes and 69 transcription factors were screened for association with SMD resistance. Six key genes, namely, *Soffic.07G0015370-1P*, *Soffic.09G0016460-1T*, *Soffic.09G0016460-1P*, *Soffic.09G0018730-3P*, *Soffic.09G0018730-3C*, and *Soffic.09G0019920-3C* and three transcription factors, namely, *Soffic.09G0015410-2T*, *Soffic.09G0017080-3C*, and *Soffic.03G0019710-2C*, were identified. These genes and transcription factors can be further explored and utilized in the marker-assisted breeding for mosaic disease resistance in sugarcane.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://www.ncbi.nlm.nih.gov/, PRJNA918436.

Author contributions

GL, YQ and LX conceptualized the study. GL, YQ and LX designed the experiments. GL, Y-BP and YQ prepared the manuscript draft. GL, Y-BP, LX, and YQ reviewed and edited the manuscript. LX provided the materials. GL, ZW, QW, WC, FX, SD and BL performed the experiments. GL and ZW conducted the data analysis. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by China Agriculture Research System of MOF and MARA [CARS-17]; Special Projects for the Central-guided Local Science and Technology Development (2022L3086); Special Fund for Science and Technology Innovation, FAFU (KFA20083A); and a Non-Funded Cooperative Agreement between the USDA-ARS and NRDCSIT on Sugarcane Breeding, Varietal Development, and Disease Diagnosis, China (Accession Number: 428234).

Acknowledgments

The unpublished reference genome of LA Purple is kindly provided by Prof. Ray Ming and Prof. Jisen Zhang from Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, China. The authors are thankful to Anna Hale, Yuling Jia, Zhongqi He, Perng-Kuang Chang for reviewing the manuscript. USDA is an equal opportunity provider and employer.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1107314/full#supplementary-material

References

Aitken, K., Farmer, A., Berkman, P., Muller, C., Wei, X., Demano, E., et al. (2017). Generation of a 345K sugarcane SNP chip. *Int. Sugar J.* 119, 816–820.

Aparicio, F., and Pallás, V. (2017). The coat protein of *Alfalfa mosaic virus* interacts and interferes with the transcriptional activity of the bHLH transcription factor ILR3 promoting salicylic acid-dependent defence signalling response. *Mol. Plant Pathol.* 18, 173–186. doi: 10.1111/mpp.12388

Asnaghi, C., Roques, D., Ruffel, S., Kaye, C., Hoarau, J., Télismart, H., et al. (2004). Targeted mapping of a sugarcane rust resistance gene (*Bru1*) using bulked segregant analysis and AFLP markers. *Theor. Appl. Genet.* 108, 759–764. doi: 10.1007/s00122-003-1487-6

Bagyalakshmi, K., Viswanathan, R., and Ravichandran, V. (2019). Impact of the viruses associated with mosaic and yellow leaf disease on varietal degeneration in sugarcane. *Phytoparasitica* 47, 591–604. doi: 10.1007/s12600-019-00747-w

Barreto, F. Z., Rosa, J. R. B. F., Balsalobre, T. W. A., Pastina, M. M., Silva, R. R., Hoffmann, H. P., et al. (2019). A genome-wide association study identified loci for yield component traits in sugarcane (*Saccharum* spp.). *PloS One* 14, e0219843. doi: 10.1371/journal.pone.0219843

Beavis, W. D. (1994). "The power and deceit of QTL experiments: lessons from comparative QTL studies," in *Proc. 49th Annu. Corn and Sorghum Int. Res. Conf.* (Washington, DC: American Seed Trade Association), 250–266.

Bello, M. H., Moghaddam, S. M., Massoudi, M., McClean, P. E., Cregan, P. B., and Miklas, P. N. (2014). Application of in silico bulked segregant analysis for rapid development of markers linked to *Bean common mosaic virus* resistance in common bean. *BMC Genomics* 15, 903. doi: 10.1186/1471-2164-15-903

Burbano, R. C. V., da Silva, M. F., Coutinho, A. E., Gonçalves, M. C., dos Anjos, I. A., Anjos, L. O. S., et al. (2022). Marker-trait association for resistance to *sugarcane mosaic virus* (SCMV) in a sugarcane (*Saccharum* spp.) panel. *Sugar Tech* 24, 1832–1844. doi: 10.1007/s12355-022-01131-5

Chang, H. H., Lee, C. H., Chang, C. J., and Jan, F. J. (2022). FKBP-type peptidyl-prolyl *cis-trans* isomerase interacts with the movement protein of *Tomato leaf curl new Delhi virus* and impacts viral replication in *Nicotiana benthamiana*. *Mol. Plant Pathol.* 23, 561–575. doi: 10.1111/mpp.13181

D'Hont, A., Ison, D., Alix, K., Roux, C., and Glaszmann, J. C. (1998). Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. *Genome* 41, 221–225. doi: 10.1139/g98-023

Da-Silva, M. F., Gonçalves, M. C., Melloni, M. N. G., Perecin, D., Landell, M. G. A., Xavier, M. A., et al. (2015a). Screening sugarcane wild accessions for resistance to *Sugarcane mosaic virus* (SCMV). *Sugar Tech* 17, 252–257. doi: 10.1007/s12355-014-0323-4

Da-Silva, M. F., Goncalves, M. C., Pinto, L. R., Perecin, D., Xavier, M. A., and Landell, M. G. A. (2015b). Evaluation of Brazilian sugarcane genotypes for resistance to sugarcane mosaic virus under greenhouse and field conditions. *Crop Prot.* 70, 15–20. doi: 10.1016/j.cropro.2015.01.002

Daugrois, J. H., Grivet, L., Roques, D., Hoarau, J. Y., Lombard, H., Glaszmann, J. C., et al. (1996). A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570'. *Theor. Appl. Genet.* 92, 1059–1064. doi: 10.1007/BF00224049

Dean, J. L. (1960). A spray method for inoculating sugarcane seedlings with the mosaic virus. *Plant Dis. Rep.* 44, 874–876.

De-Souza, I., Macêdo, G. A. R., Barbosa, M. H. P., Barros, B. D. A., Carvalho, S. G. M., and Xavier, A. D. S. (2017). Reaction of sugarcane genotypes to strains of the *Sugarcane mosaic virus*. *Int. J. Curr. Res.* 9, 59112–59119.

Duble, C. M., Melchinger, A. E., Kuntze, L., Stork, A., and Lubberstedt, T. (2000). Molecular mapping and gene action of *Scm1* and *Scm2*, two major QTL contributing to SCMV resistance in maize. *Plant Breed.* 119, 299–303. doi: 10.1046/j.1439-0523.2000.00509.x

Dussle, C. M., Quint, M., Melchinger, A. E., and Lübberstedt, T. (2003). Saturation of two chromosome regions conferring resistance to SCMV with SSR and AFLP markers by targeted BSA. *Theor. Appl. Genet.* 106, 485–493. doi: 10.1007/s00122-002-1107-x

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* 9, 275–296. doi: 10.1146/annurev.py.09.090171.001423

Grisham, M. P. (2011). "Mosaic," in *A guide to sugarcane diseases*. Eds. P. Rott, R. A. Bailey, J. C. Comstock and B. J. Croft (Montepellier: CIRAD Publication Services), 249–254.

Gu, K., Zhang, Q., Yu, J., Wang, J., Zhang, F., Wang, C., et al. (2020). R2R3-MYB transcription factor *MdMYB73* confers increased resistance to the fungal pathogen *Botryosphaeria dothidea* in apples *via* the salicylic acid pathway. *J. Agr. Food Chem.* 69, 447–458. doi: 10.1021/acs.jafc.0c06740

Jia, Y., McAdams, S. A., Bryan, G. T., Hershey, H., and Valent, B. (2000). Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. *EMBO J.* 19, 4004–4014.

Jones, J. D. G., and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323–329. doi: 10.1038/nature05286

Koike, H., and Gillaspie, J. R. (1989). "Mosaic," in *Disease of sugarcane: Major disease*. Eds. C. Ricaud, B. T. Egan and A. G. Gillaspie (Amsterdam: Elsevier Science Publisher), 301–322.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res.* 37, 4181–4193. doi: 10.1093/nar/gkp552

Lavín-Castaeda, J., Sentíes-Herrera, H. E., Trejo-Téllez, L. I., Bello-Bello, J. J., and Gómez-Merino, F. C. (2020). Advances in the selection program of sugarcane (*Saccharum* spp.) varieties in the colegio de postgraduados. *AGRO Productividad* 13, 123–129. doi: 10.32854/agrop.v13i11.1776

Lee, K. J., and Kim, K. (2015). The rice serine/threonine protein kinase *OsPBL1* (*ORYZA SATIVA ARABIDOPSIS PBS1-LIKE 1*) is potentially involved in resistance to rice stripe disease. *Plant Growth Regul.* 77, 67–75. doi: 10.1007/s10725-015-0036-z

Li, Q., Chen, Z., and Liang, H. (2000). *Modern sugarcane improvement technology* (Guangzhou: South China University of Technology Press), 46–47.

Li, W., Deng, Y., Ning, Y., He, Z., and Wang, G. L. (2020). Exploiting broad-spectrum disease resistance in crops: from molecular dissection to breeding. *Annu. Rev. Plant Biol.* 71, 575–603. doi: 10.1146/annurev-arplant-010720-022215

Li, W., Shan, H., Zhang, R., Wang, X., Luo, Z., Yin, J., et al. (2018). Screening for resistance to *Sugarcane streak mosaic virus* and *Sorghum mosaic virus* in new elite sugarcane varieties/clones from China. *Acta Phytopathol. Sin.* 48, 389–394. doi: 10.13926/j.cnki.apps.000220

Liu, S., Yang, X., Zhang, D., Bai, G., Chao, S., and Bockus, W. (2014). Genome-wide association analysis identified SNPs closely linked to a gene resistant to *Soil-borne wheat mosaic virus*. *Theor. Appl. Genet.* 127, 1039–1047. doi: 10.1007/s00122-014-2277-z

Li, W., Wang, X., Huang, Y., Shan, H., Luo, Z., Ying, X., et al. (2013). Screening sugarcane germplasm resistant to *Sorghum mosaic virus*. *Crop Prot.* 43, 27–30. doi: 10.1016/j.cropro.2012.09.005

Lu, G., Wang, Z., Xu, F., Pan, Y. B., Grisham, M. P., and Xu, L. (2021). Sugarcane mosaic disease: characteristics, identification and control. *Microorganisms* 9, 1984. doi: 10.3390/microorganisms9091984

McCouch, S. R., Cho, Y. G., Yano, M., Paul, E., Blinstrub, M., Morishima, H., et al. (1997). Report on QTL nomenclature. *Rice Genet. Newsl.* 14, 11–13.

Noman, A., Hussain, A., Adnan, I., Ashraf, M. F., Zaynab, M., and Khan, K. A. (2019). A novel MYB transcription factor *CaPHL8* provide clues about evolution of pepper immunity against soil borne pathogen. *Microb. Pathogenesis* 137, 103758. doi: 10.1016/j.micpath.2019.103758

O'Conner, S., Zheng, W., Qi, M., Kandel, Y., Fuller, R., Whitham, S. A., et al. (2021). *GmNF-YC4-2* increases protein, exhibits broad disease resistance and expedites maturity in soybean. *Int. J. Mol. Sci.* 22, 3586. doi: 10.3390/ijms22073586

Osuna-Cruz, C. M., Paytuvi-Gallart, A., Di Donato, A., Sundesha, V., Andolfo, G., Aiese Cigliano, R., et al. (2018). PRGdb 3.0: a comprehensive platform for prediction and analysis of plant disease resistance genes. *Nucleic Acids Res.* 46, D1197–D1201. doi: 10.1093/nar/gkx1119

Poland, J. A., Balint-Kurti, P. J., Wisser, R. J., Pratt, R. C., and Nelson, R. J. (2009). Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci.* 14, 21–29. doi: 10.1016/j.tplants.2008.10.006

Qi, D., Dubiella, U., Kim, S. H., Sloss, D. I., Dowen, R. H., Dixon, J. E., et al. (2014). Recognition of the protein kinase AVRPPHB SUSCEPTIBLE1 by the disease resistance protein RESISTANCE TO PSEUDOMONAS SYRINGAE5 is dependent on *s*-acylation and an exposed loop in AVRPPHB SUSCEPTIBLE1. *Plant Physiol.* 164, 340–351. doi: 10.1104/pp.113.227686

Raboin, L. M., Oliveira, K. M., Lecunff, L., Telismart, H., Roques, D., Butterfield, M., et al. (2006). Genetic mapping in sugarcane, a high polyploid, using bi-parental progeny: identification of a gene controlling stalk colour and a new rust resistance gene. *Theor. Appl. Genet.* 112, 1382–1391. doi: 10.1007/s00122-006-0240-3

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5, 94–100. doi: 10.1016/S1369-5266(02)00240-6

Roach, B. T. (1989). "Origin and improvement of the genetic base of sugarcane," in *Proc. aust. Soc. sugarcane techno*. Ed. B. T. Egan (Tweed Heads, Australia: Annual Conference), 10, 34–47.

Roossinck, M. J. (2015). Plants, viruses and the environment: ecology and mutualism. *Virology* 479, 271–277. doi: 10.1016/j.virol.2015.03.041

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Shan, H., Huan, Y., Wang, X., Li, J., Zhang, R., Cang, X., et al. (2020). One-step multiplex RT-PCR detection technique for simultaneous detection of three pathogens of sugarcane mosaic disease. Patent CN110951924A.

Singh, R. K., Banerjee, N., Khan, M. S., Yadav, S., Kumar, S., Duttamajumder, S. K., et al. (2016). Identification of putative candidate genes for red rot resistance in sugarcane (*Saccharum* species hybrid) using LD-based association mapping. *Mol. Genet. Genomics* 291, 1363–1377. doi: 10.1007/s00438-016-1190-3

Singh, V., Sinha, O. K., and Kumar, R. (2003). Progressive decline in yield and quality of sugarcane due to *Sugarcane mosaic virus*. *Indian Phytopathol.* 56, 500–502.

Smakowska-Luzan, E., Mott, G. A., Parys, K., Stegmann, M., Howton, T. C., Layeghifard, M., et al. (2018). An extracellular network of *Arabidopsis* leucine-rich repeat receptor kinases. *Nature* 553, 342–346. doi: 10.1038/nature25184

Viswanathan, R., and Balamuralikrishnan, M. (2005). Impact of mosaic infection on growth and yield of sugarcane. *Sugar Tech* 7, 61–65. doi: 10.1007/BF02942419

Wang, Z., Lu, G., Wu, Q., Li, A., Que, Y., and Xu, L. (2021). Isolating QTL controlling sugarcane leaf blight resistance using a two-way pseudo-testcross strategy. *Crop J.* 10, 1131–1140.

Wu, L., Zu, X., Wang, S., and Chen, Y. (2012). Sugarcane mosaic virus-long history but still a threat to industry. *Crop Prot.* 42, 74–78. doi: 10.1016/j.cropro.2012.07.005

Xia, X., Melchinger, A. E., Kuntze, L., and Lübberstedt, T. (1999). Quantitative trait loci mapping of resistance to *sugarcane mosaic virus* in maize. *Phytopathology* 89, 660–667. doi: 10.1094/PHYTO.1999.89.8.660

Xu, M., Melchinger, A. E., and Lübberstedt, T. (2000). Origin of *Scm1* and *Scm2* - two loci conferring resistance to *sugarcane mosaic virus* (SCMV) in maize. *Theor. Appl. Genet.* 100, 934–941. doi: 10.1007/s001220051373

Yakupjan, H., Asigul, I., Wang, Y., and Liu, Y. (2015). Advances in genetic engineering of plant virus resistance. *Chin. J. Biotechnol.* 31, 976–994. doi: 10.13345/j.cjb.150042

Yang, C. (2015). *Construction of high-density genetic map and location of QTLs for smut in saccharum spontaneum l.. Ph.D. thesis* (Nanning: Guangxi University).

Yang, Q., He, Y., Kabahuma, M., Chaya, T., Kelly, A., Borrego, E., et al. (2017). A gene encoding maize caffeoyl-CoA *O*-methyltransferase confers quantitative resistance to multiple pathogens. *Nat. Genet.* 49, 1364–1372. doi: 10.1038/ng.3919

Yang, X., Islam, M. S., Sood, S., Maya, S., Hanson, E. A., Comstock, J., et al. (2018). Identifying quantitative trait loci (QTLs) and developing diagnostic markers linked to orange rust resistance in sugarcane (*Saccharum* spp.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00350

Yang, X., Sood, S., Glynn, N., Islam, M., Comstock, J., and Wang, J. (2017). Constructing high-density genetic maps for polyploid sugarcane (*Saccharum* spp.) and identifying quantitative trait loci controlling brown rust resistance. *Mol. Breed.* 37, 116. doi: 10.1007/s11032-017-0716-7

Yang, R., Zhou, H., Xiao, W., Lü, D., Liao, H. X., Chen, D. D., et al. (2020). Testing on sugarcane mosaic resistance of sugarcane major parents under field conditions. *Sugar Crop China* 42, 47–52. doi: 10.13570/j.cnki.scc.2020.02.009

You, Q., Sood, S., Luo, Z., Liu, H., Islam, M. S., Zhang, M., et al. (2020). Identifying genomic regions controlling ratoon stunting disease resistance in sugarcane (*Saccharum* spp.) clonal F_1 population. *Crop J.* 9, 1070–1078. doi: 10.1016/j.cj.2020.10.010

You, Q., Yang, X., Peng, Z., Islam, M. S., Sood, S., Luo, Z., et al. (2019). Development of an axiom Sugarcane100K SNP array for genetic map construction and QTL identification. *Theor. Appl. Genet.* 132, 2829–2845. doi: 10.1007/s00122-019-03391-4

Yuan, L., Duble, C. M., Muminovic, J., Melchinger, A. E., and Lübberstedt, T. (2004). Targeted BSA mapping of *Scmv1* and *Scmv2* conferring resistance to SCMV using *Pst*I/*Mse*I compared with *Eco*RI/*Mse*I AFLP markers. *Plant Breed.* 123, 434–437. doi: 10.1111/j.1439-0523.2004.00966.x

Zhang, H., Chen, C., Li, L., Tan, X., Wei, Z., Li, Y., et al. (2021). A rice LRR receptor-like protein associates with its adaptor kinase OsSOBIR1 to mediate plant immunity against viral infection. *Plant Biotechnol. J.* 19, 2319–2332. doi: 10.1111/pbi.13663

Zhang, G., Chen, M., Li, L., Xu, Z., Chen, X., Guo, J., et al. (2009). Overexpression of the soybean *GmERF3* gene, an AP2/ERF type transcription factor for increased tolerances to salt, drought, and diseases in transgenic tobacco. *J. Exp. Bot.* 60, 3781–3796. doi: 10.1093/jxb/erp214

Zhang, S., Li, X., Wang, Z., George, M. L. C., Jeffers, D. P., Wang, F., et al. (2003). QTL mapping for resistance to SCMV in Chinese maize germplasm. *Maydica* 48, 307–312.

Zhou, F. (2015). *Molecular detection and physiological and biochemical changes of sugarcane mosaic disease. master's thesis* (Nanning: Guangxi University).

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/fpls.2023.1131975

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Youlu Yuan,
Institute of Cotton Research (CAAS), China
Luming Yang,
Henan Agricultural University, China

*CORRESPONDENCE
Zhonghua Wang
zhonghuawang@nwafu.edu.cn
Hongxian Mei
meihx2003@126.com
Bing Jing
jingbing@nwafu.edu.cn

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 26 December 2022
ACCEPTED 08 February 2023
PUBLISHED 23 February 2023

CITATION
Wang H, Cui C, Liu Y, Zheng Y, Zhao Y, Chen X, Wang X, Jing B, Mei H and Wang Z (2023) Genetic mapping of QTLs controlling brown seed coat traits by genome resequencing in sesame (*Sesamum indicum* L.). *Front. Plant Sci.* 14:1131975. doi: 10.3389/fpls.2023.1131975

Genetic mapping of QTLs controlling brown seed coat traits by genome resequencing in sesame (*Sesamum indicum* L.)

Han Wang[1†], Chengqi Cui[2,3†], Yanyang Liu[2,3], Yongzhan Zheng[2,3], Yiqing Zhao[1], Xiaoqin Chen[1], Xueqi Wang[1], Bing Jing[1*], Hongxian Mei[2,3*] and Zhonghua Wang[1*]

[1]State Key Laboratory of Crop Stress Biology for Arid Areas, College of Agronomy, Northwest A&F University, Yangling, China, [2]Henan Sesame Research Center, Henan Academy of Agricultural Sciences, Zhengzhou, China, [3]The Shennong Laboratory, Zhengzhou, China

Introduction: Sesame seeds have become an irreplaceable source of edible oils and food products with rich nutrients and a unique flavor, and their metabolite contents and physiological functions vary widely across different seed coat colors. Although the quantitative trait loci (QTLs) for genetic variation in seed coat color have been extensively investigated, the identification of unique genetic loci for intermediate colors such as brown has not been reported due to their complexity.

Methods: Here, we crossed the white sesame 'Yuzhi No. 8' (YZ8) and the brown sesame 'Yanzhou Erhongpi' (YZEHP) to construct a recombinant inbred line (RIL) population with consecutive self-fertilization for ten generations.

Results: The selfed F1 seeds were brown which was controlled by a dominant gene. Based on the genotyping by whole-genome resequencing of the RILs, a major-effect QTL for brown coat color was identified through both bulk segregant analysis (BSA) and genetic linkage mapping in sesame, which was located within a 1.19 Mb interval on chromosome 6 (qBSCchr6). Moreover, we found that the YZEHP seed coat initially became pigmented at 20 days post-anthesis (DPA) and was substantially colored at 30 DPA. We screened 13 possible candidate genes based on the effects of genetic variants on protein coding and predicted gene functions. Furthermore, qRT–PCR was used to verify the expression patterns of these genes in different post-anthesis developmental periods. We noted that in comparison to YZ8 seeds, YZEHP seeds had expression of SIN_1023239 that was significantly up-regulated 2.5-, 9.41-, 6.0-, and 5.9-fold at 15, 20, 25, and 30 DPA, respectively, which was consistent with the pattern of brown seed coat pigment accumulation.

Discussion: This study identified the first major-effect QTL for the control of the brown seed coat trait in sesame. This finding lays the foundation for further fine mapping and cloning as well as investigating the regulatory mechanism of seed coat color in sesame.

KEYWORDS

sesame, seed coat color, whole-genome resequencing, BSA, QTL mapping, qRT-PCR

1 Introduction

Sesame (*Sesamum indicum* L.) is an exceptional and essential oilseed crop; it is one of the oldest such crops known to mankind, having been domesticated from its wild progenitor *S. malabaricum* on the Indian subcontinent approximately 5000 years ago (Bedigian, 2003; Fuller, 2003). Sesame seeds are used for a wide variety of applications, both as condiments and as a source of edible oil. Sesame oil is commonly used for its distinctive flavor, in addition to being a key component in the production of margarine, soap, and lubricants (Hwang, 2005). One of the main distinguishing characteristics of sesame seeds is the color of the seed coat. Seed coat color is a crucial aspect of seed quality and is related to the biochemical properties of the seed and to the activity and content of its antioxidant substances (Shahidi et al., 2006; Kermani et al., 2019). These different biochemical and antioxidant properties may be most closely related to higher levels of sesamol, sesaminol, alpha-tocopherol, and flavonoids in the seed coats of colored sesame than that of white sesame seeds (Xu et al., 2005). However, it has not yet been possible to identify the genes that regulate the metabolic pathways and mechanisms of interaction that determine sesame seed coat color, which is typically thought to show a complicated pattern of quantitative inheritance (Zhang et al., 2013).

Mature sesame seeds come in a variety of natural coat colors, including black, gray, brown, gold, yellow, beige, and white (Prasad and Gangopadhyay, 2011; Pandey et al., 2013). As seed coat color is one of the central targets of sesame breeding programs, research into the inheritance of the trait and the corresponding gene loci have been of considerable scientific interest. In 1931, a Japanese researcher initially suggested that the inheritance of sesame seed coat color potentially fit a segregation pattern involving three allelic genes (Teshima, 1931). Zhang et al. (2013) identified and analyzed the genetic segregation of quantitative trait loci (QTLs) for sesame seed coat color over six generations and concluded that two major-effect genes with additive-dominant-epistatic effects and multiple minor-effect genes with additive-dominant-epistatic effects were responsible for controlling the seed coat color trait. Moreover, seven QTLs that control sesame seed coat color traits were identified by Du et al. (2019). In addition, Wang et al. (2016) mapped three QTLs that were repeatedly detected and accounted for 80% of the phenotype variation by resequencing a RIL population. According to the annotation of genes anchored to genomic intervals combined with transcriptome analysis, the polyphenol oxidase (PPO) gene may be involved in the production of the black seed coat in sesame, and this finding has been supported by several investigations (Wei et al., 2015; Wang et al., 2016; Wei et al., 2016; Wang et al., 2020). Furthermore, since the development of next-generation sequencing technologies, whole-genome association analysis has been used to dissect complex traits in crops, as QTL mapping research in the segregating progeny of classical hybrids is limited by a low number of recombination events and cultivar-specific allelic loci (Nordnorg and Welgel, 2008; Guo et al., 2013). By resequencing an association analysis panel of 366 sesame germplasm lines, Cui et al. (2021) demonstrated complex genetic variation in seed coat color. The results revealed that 22 significant single-nucleotide polymorphisms (SNPs) were located within the reported QTL confidence intervals and that the four most reliable and significant flanking regions of these SNPs contained 92 candidate genes. However, researchers have been unable to perform additional in-depth investigations on the locus that controls the seed coat color trait in sesame due to gaps in the QTL mapping studies regarding intermediate seed coat colors. Furthermore, it is not possible to validate the currently available genetic loci against each other because much of the existing sesame QTL mapping research has been based on independent genetic maps. Thus, to meet the needs of molecular breeding, QTL mapping research on sesame seed coat color should be expanded using high-quality genomes anchored to chromosomes.

Plant seed color is mainly characterized by the accumulation of pigmented metabolites in the seed coat. In this context, a brown seed coat has been identified as possibly being regulated by the flavonoid synthesis pathway in several plant species. The genes that may regulate the brown seed coat in *Arabidopsis* include those encoding the Transparent Testa12 (TT12) and EXO70 exocyst subunit (EXO70B1) transporter proteins and the proanthocyanidin (PA) oxidase enzyme (TT10) (Debeaujon et al., 2001; Pourcel et al., 2005; Kulich et al., 2013). Moreover, Transparent Testa Glabra2 (TTG2) was found to interact with TTG1 to form a complex that directly regulates the expression of TT12 to produce brown *Arabidopsis* seed coats (Gonzalez et al., 2016). Among other crops, many transcription factors, such as MYB, basic helix-loop-helix (bHLH), and WD40 proteins, have been identified as potentially being involved in the regulation of flavonoid biosynthesis (Zhang et al., 2009; Gillman et al., 2011; Hong et al., 2017; Ren et al., 2017). Small interfering RNAs (siRNAs) were also found to silence the expression of transposable elements (TEs) or protein-coding genes and thereby affect the synthesis and regulation of flavonoid metabolites (Jia et al., 2020). In addition, PPOs such as laccase, tyrosinase, and even peroxidase are involved in the oxidation steps of PA, lignin, and melanin biosynthesis (Pourcel et al., 2007; Yu, 2013).

In this study, we used a RIL population and the whole-genome resequencing technique to perform QTL mapping for seed coat color in sesame. A major-effect QTL, *qBSCchr6*, controlling the brown seed coat trait in sesame was revealed by the combination of BSA and high-density genetic linkage mapping. The candidate genes involved in the regulation of the brown seed coat were screened based on the evaluation of the effect of genetic variants on protein coding and predicted gene functions. The expression patterns of these genes in different developmental periods at post-anthesis were analyzed using qRT–PCR. The results of this study will enhance the development of research on the genetic and molecular mechanisms of sesame seed coat color regulation and provide a basis for functional gene cloning studies.

2 Materials and methods

2.1 Plant materials

The cultivar Yanzhou Erhongpi (YZEHP) has a brown seed coat and is a landrace collected from Shandong Province, China. The Yuzhi No. 8 (YZ8) cultivar, which was bred by Henan Academy of Agricultural Science, produces seeds with a white coat. A mapping

population of 315 recombinant inbred lines (RILs, F_{10} generation) was constructed from a cross between YZEHP and YZ8 using the single-seed descent (SSD) method. The lines showed obvious differences in traits such as plant height, thousand grain weight, capsule length, and seed coat color. The RIL population and both parents were planted in 2020 at experimental sites in Sanya, Hainan Province (SY, N18°140′, E109°290′), Zhumadian, Henan Province (ZMD, N32°59′, E114°42′), and Nanyang, Henan Province (NY, N32°54′, E112°24′). All the plants were arranged in a randomized block design with two replicates, and 10 representative plants of each line were harvested for the investigation of seed coat color.

2.2 Seed coat color evaluation and statistical analysis

Initially, we superficially observed both brown and white mature seed coat types. Additionally, a Colorflex EZ spectrophotometer (Hunter Associates Laboratory Inc, Virginia, USA) was used to measure the colors of the seed coats in three different environments. Mature seeds were scanned in a quartz box to quantify the L*, a*, and b* values for seed coat color. The L* value, which represents brightness, ranges from 0 (black) to 100 (white), while the values of a* and b*, which represent color shades, range from -60 for green to +60 for red and -60 for blue to +60 for yellow, respectively (Aruldass et al., 2014). Phenotypic statistics were calculated using SAS v9.1 (SAS Institute, Inc., Cary, NC, USA). Based on the mean values of L*, a*, and b* for the sesame seed coat color phenotype among replicates and different environments, the broad-sense heritability was calculated using the AOV module in QTL IciMapping v4.2 (Meng et al., 2015). Furthermore, the color phenotypes observed for each line corresponded to the L*, a*, and b* values and were visualized by ggplot2 v3.3.6 (Wickham, 2016).

2.3 Sequencing and SNP/InDel calling

Genomic DNA was extracted from seedling leaves of the parents and RILs using a modified cetyltrimethylammonium bromide (CTAB) method (Mei et al., 2017). The quality of the genomic DNA was examined with a NanoDrop 2000 (Thermo Fisher Scientific, Waltham, MA, USA) and by 1.0% agarose gel electrophoresis. After ultrasound fracturing, the DNA was sequentially end repaired, sequencing junction ligated, and enriched by magnetic bead adsorption to obtain fragments with a genomic length of approximately 400 bp. These fragments were then amplified by PCR to establish a sequencing library. The Illumina NovaSeq 6000 platform was used to sequence the quality-checked libraries with a total sequencing read length of 300 bp using the Illumina PE150 sequencing strategy. The two parents and the RILs were sequenced at depths of approximately 15× and 5×, respectively. The reads were filtered to eliminate adapters and low-quality reads. Based on the seed coat color phenotypes of the RILs grown in ZMD, we merged the clean reads of 50 randomly selected lines from white and brown sesame, respectively, to construct extreme bulks. The clean reads of all samples were aligned to the reference genome (Wang et al., 2016) using Burrows–Wheeler Aligner (BWA) v0.7.17 (Li and Durbin, 2009). SNPEff v4.3T (Cingolani et al., 2012) and the gene annotation information of the reference genome were used to functionally annotate SNPs and small InDels after correction and detection by using Genome Analysis Toolkit (GATK) v4.0.11.0 (McKenna et al., 2010) and SAMtools v1.9.0 (Li et al., 2009). According to genetic principles, all markers were examined for parental polymorphism. Variant sites that differed between the parents were selected and coded as molecular markers, and the genotypes of the RILs and bulks were extracted for additional analysis.

2.4 BSA, genetic map construction, and QTL mapping

The QTL-seq method was implemented to calculate the ΔSNP index (Takagi et al., 2013). The SNP index represents the proportion of short reads harboring SNPs that differ from the reference sequence to the total reads covering a particular genomic position (Abe et al., 2012). The SNP index of the extreme bulks was statistically analyzed based on the average SNP index within each genomic interval containing 20 SNP variants, which was individually measured using a sliding window of 5 SNP variants. The ΔSNP index is the average SNP index difference between the two extreme bulks (99.9% quantile as the threshold), and this analysis revealed significant differences in genotype frequencies between the extreme bulks (Hill et al., 2013).

We selected polymorphic markers of the aa×bb type between the parents as valid markers, and these markers were screened for abnormal bases, completeness, and segregation distortion after being used to genotype the RIL population. Moreover, we utilized a reference genome assisted correction-based linkage group ordering scheme. We completed the construction of the genetic map using MstMap (Wu et al., 2008), and we then used ASMapR v1.0-4 and R/qtl v1.44-9 to evaluate the monomeric origin and recombination relationships (Broman et al., 2003; Taylor and Butler, 2017). In addition, we analyzed the collinearity of the linkage map with the physical map. Finally, the visualization of the genetic map was completed using LinkageMapViewR v2.1.2 (Ouellette et al., 2018). R/qtl was used for standard and stepwise interval mapping with 1000 permutations and a p value of 0.05 as the logarithm of odds (LOD) significance detection threshold. Composite interval mapping (CIM) was performed based on a 5 cM marker window size and a step of 1 cM. The location of each QTL was determined based on the location of the LOD peak for each QTL and the surrounding area. The percentage of phenotypic variation explained (R^2) by the QTL was estimated at the highest probability peak (Tao et al., 2022).

2.5 Bioinformatic analysis

Gene sequence information was obtained based on the candidate intervals. The functions of the candidate genes were annotated by using the NR (http://www.ncbi.nlm.nih.gov/), UniProt (http://www.uniprot.org/), Gene Ontology (GO) (http://www.geneontology.org/), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) databases, and

the Basic Local Alignment Search Tool (BLAST) program in the EggNOG (http://eggnog-mapper.embl.de/) database for annotation. Moreover, the analysis of protein coding variants included variants annotated by SNPEff with sequence ontology terms for assessing sequence changes and impacts, and categorized the impact of SNP/InDel within the candidate interval into four classes: High, Moderate, Low, and Modifier, in descending order according to the effect of the variant on protein coding (Supplementary Table 1) (Cingolani et al., 2012; Oren et al., 2022).

2.6 RNA extraction and qRT–PCR analysis of candidate genes

We also sampled parental seeds at 10, 15, 20, 25, and 30 DPA in Yangling, Shaanxi Province (N34°27′, E108°07′), in 2022. Quantitative color analysis of the seed coat was performed with a CIE-Lab color scale (Colorimeter, CS-820, Hangzhou, China) with a 6 mm aperture due to the small sample size (Dong et al., 2022). All samples were flash frozen in liquid nitrogen and stored at -80°C in the refrigerator until needed. Total seed RNA was extracted using a kit (DP441, TIANGEN, China) and first-strand cDNA was synthesized by the PrimeScript RT reagent kit (#6210A, Takara, Kusatsu, Japan). Three independent biological replicates of the qRT–PCR (#RR820A, Takara, Kusatsu, Japan) protocol were tested using cDNA as the template for each experiment. The sesame actin gene (SIN_1006268) was used as the internal reference gene (Wei et al., 2015), and relative gene expression was calculated using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001).

3 Results

3.1 Phenotypic and genetic analysis of the brown seed coat in sesame

To reveal the genetic basis of the brown seed coat color in sesame, a RIL population including 315 lines was developed using YZEHP (male, brown seeds) and YZ8 (female, white seeds) as two parental lines in this study. We first investigated the phenotypes of seed coat color traits for several generations. The contemporary hybrid seeds obtained from the maternal plants were white (consistent with the YZ8 phenotype), while the selfed seeds developed from the F_1 generation were brown (consistent with YZEHP phenotype) (Figure 1A). The brown seed coat in sesame is dominant to the white seed coat. Notably, angiosperm seed coats develop from bead tepals (Haughn and Chaudhury, 2005). Therefore, the genotype of the sesame seed coat is consistent with that of the female parent because the inheritance of sesame seed coat traits is matrilineal, as found in previous studies (Wang et al., 2016; Das et al., 2018). We performed visual observations of mature seed color phenotypes and identified 162 and 153 lines among 315 RILs with brown and white seed coats, respectively (data not shown). Furthermore, we quantified the seed coat color by using a colorimeter and found that the L*, a*, and b* values of the brown and white seeds of the RILs differed significantly (P<0.001) across the three environments (Figure 1B). Interestingly, the L*, a*, and b* values showed a bimodal continuous distribution in the RIL population (Supplementary Figure 1). Additionally, the mean coefficients of variation (CV) for the L*, a*, and b* values across environments were 6.54%, 27.86%, and 12.80%, respectively. The L* value for RILs across environments ranged from 49.23~64.63, the a* value ranged from 4.51~11.18, and the b* value ranged from 18.36~28.97. The L*, a*, and b* values presented average broad-sense heritabilities of 94.95%, 96.87%, and 95.67%, respectively (Figure 1B; Table 1). The results suggest that the phenotype of the brown seed coat trait in sesame is determined (in order from highest to lowest) by redness, yellowness, and brightness.

3.2 Sequencing the RIL population for BSA analysis and marker identification

Whole-genome resequencing was used to analyze the two parents and 315 RILs. A total of 455.90 Gb of clean bases was obtained after sequencing and filtering; the average Q30 quality score was over 90.98%; the average matching efficiency of the samples to reference genome was 97.18%; and the GC content

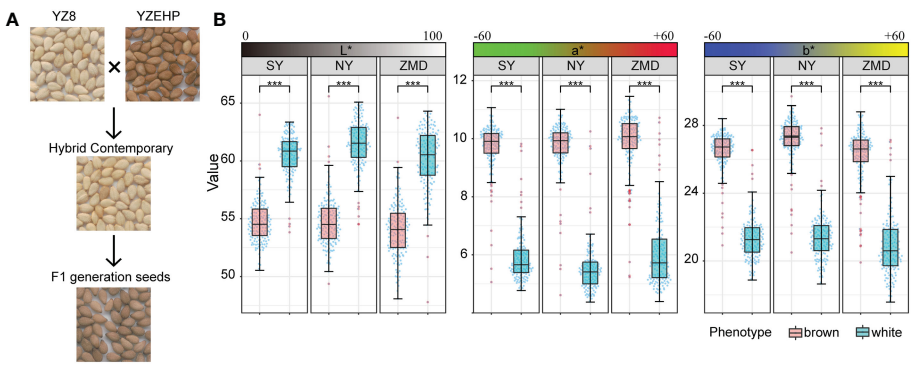


FIGURE 1
Phenotypic analysis of parents and RILs. **(A)** Seed coat color of the parents and hybrid offspring. **(B)** Distribution of quantitative values of L*, a*, and b* in the RIL population, Significant levels were determined by T-test, with *** representing p<0.001 level.

TABLE 1 Descriptive statistics and broad-sense heritability (H^2) for three seed coat color related traits of RILs.

Trait	Environment	Mean	SD	Range	CV (%)	Excess Kurtosis	Skewness	H^2 (%)
L*	SY	57.51	3.41	50.54-63.99	5.94	-1.33	0.04	94.95
	NY	58.01	3.95	49.36-65.59	6.82	-1.32	0.06	
	ZMD	57.01	3.92	47.79-64.3	6.88	-1.04	0.01	
	Mean	57.51	3.76	49.23-64.63	6.54	-1.23	0.04	
a*	SY	7.86	2.09	4.77-11.07	26.55	-1.76	-0.07	96.87
	NY	7.67	2.28	4.37-11.01	29.78	-1.81	-0.04	
	ZMD	8.04	2.21	4.39-11.46	27.49	-1.65	-0.14	
	Mean	7.86	2.19	4.51-11.18	27.86	-1.74	-0.08	
b*	SY	23.99	2.85	18.87-28.40	11.87	-1.63	-0.12	95.67
	NY	24.34	3.20	18.64-29.72	13.15	-1.65	-0.09	
	ZMD	23.64	3.16	17.57-28.80	13.35	-1.49	-0.14	
	Mean	23.99	3.07	18.36-28.97	12.80	-1.59	-0.12	

SD, standard deviation; CV, coefficient of variation; H^2 broad-sense heritability.

ranged from 36.67~39.3%. The amounts of data obtained for YZEHP and YZ8 were 5.24 Gb and 4.93 Gb, respectively, and the actual average amount of data obtained for the RILs was 1.41 Gb, and the average sequencing coverage was 18.61× for the parents and 5.16× for the RIL population (Supplementary Figure 2; Supplementary Table 2). It was evident that all samples showed a sufficient amount of data, normal distribution, and regular sequencing results when compared to the sesame reference genome, suggesting that they could be used for subsequent analysis. Then, we merged the clean reads separately from 50 lines to develop the following two extreme bulks: one with 231 million reads in a white seed coat bulk and the other with 240 million reads in a brown seed coat bulk (Supplementary Table 2). These two extreme bulks were screened for 38,752 SNP markers, which were used to calculate genotype frequencies (Supplementary Table 3). Additionally, 1,284,658 SNP/InDel markers were detected between two parental lines, of which 167,862 were valid markers of the aa×bb type with a sequencing depth of no less than 2 in the RILs and 10 in the parental lines (Supplementary Figure 3). After screening the markers for abnormal bases, completeness, and segregation distortion, 7,908 high-quality markers remained after genotyping the RIL population with validated polymorphic markers were used for the following analysis.

3.3 Construction of a high-density genetic map

Among the remaining 7,908 markers, 7,817 were ordered into 13 linkage groups based on the reference genome. The length of the high-density linkage map was 1833.89 cM, and the average distance between markers was 0.23 cM (Supplementary Figure 4; Table 2; Supplementary Table 4). The linkage group with the highest number of markers was LG5, which contained 1,667 markers. We next performed a quality assessment analysis of the genetic map. First, based on haplotype map analysis of recombination breakpoints, 7,817 markers were used to genotype the RILs, and the sources of recombination blocks were specifically explained (Supplementary Figure 5). Second, we analyzed the relationships between the positions of all mapped markers in the genetic map and the physical map of the reference genome, and the Spearman correlation coefficient between them exceeded 0.89, with a high observed collinearity (Supplementary Figure 6; Supplementary Table 5). Third, we used a heatmap to directly reflect recombination rates and LOD scores between markers, and no switched alleles were discovered; switched alleles were indicated by low LOD scores and low recombination fractions (Supplementary Figure 7) (Maldonado-Taipe et al., 2022). In summary, we constructed an accurate and reliable genetic map which was suitable for QTL mapping.

3.4 BSA and QTL mapping reveal the physical position of the locus controlling the brown seed coat in sesame

We identified QTLs using both BSA and traditional linkage mapping methods. In BSA, the SNP index of the two extreme bulks was calculated and visualized using sliding window analysis along chromosomes. Based on a 99.9% quantile threshold, we identified a significant physical interval (16.36 Mb~21.46 Mb) on chr6 by analyzing the SNP index of the two bulks throughout the 38,752 SNP markers (Figure 2A). In particular, the mean SNP index of the two bulks within the 18,323,068 to 20,213,179 bp sliding window was 0.89 and 0.14, respectively (Supplementary Table 6). This result suggests that there was a strong signal in this genomic region which may be controlled by a powerful QTL. To map brown seed coat-related QTLs more accurately, linkage mapping was performed based on the high-density genetic map and quantitative data for RILs seed coat color. We examined QTLs in three environments for L*, a*, and b* values.

10.3389/fpls.2023.1131975

TABLE 2 Basic information of the high-density genetic linkage map of RIL population.

Linkage group ID	Total marker	Total distance (cM)	Average distance (cM)	Max gap (cM)	Gaps < 5cM (%)
LG1	333	125.16	0.38	6.20	98.50
LG2	340	155.93	0.46	13.84	97.94
LG3	691	153.39	0.22	7.04	99.42
LG4	533	146.11	0.27	9.34	98.31
LG5	1667	142.25	0.09	12.18	99.94
LG6	881	125.09	0.14	8.40	99.55
LG7	817	143.49	0.18	7.72	99.14
LG8	508	164.67	0.32	11.47	98.62
LG9	402	121.01	0.30	9.74	99.00
LG10	784	127.72	0.16	8.53	99.74
LG11	297	121.37	0.41	11.41	97.64
LG12	527	143.52	0.27	10.41	99.43
LG13	37	164.19	4.56	18.37	67.57
Total	7817	1833.89	0.23	18.37	99.08

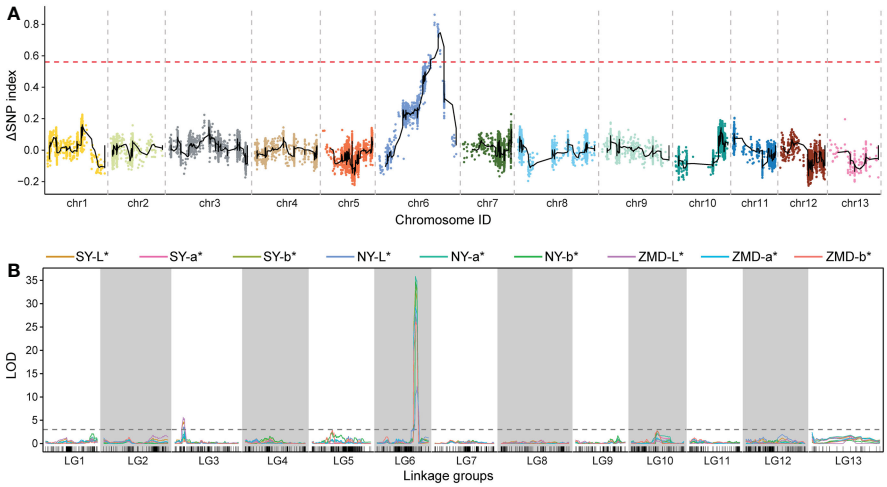


FIGURE 2
QTL identification across chromosomes and linkage groups using BSA and genetic linkage mapping, respectively. **(A)** QTL-seq analysis with the number of SNPs as the sliding window, with the red dashed line representing the significance threshold. **(B)** QTL scanning of the brown seed coat for the total linkage groups.

Under the threshold condition of LOD≥3.10 (p value=0.05), three major QTLs were detected in all three environments within a genetic interval of 89.17~101.29 cM on chr6 (Figure 2B; Table 3). The mean LOD values of the QTLs for L*, a*, and b* in the three environments were 24.27, 33.02, and 31.51, respectively, and the mean R^2 were 33.64%, 36.63%, and 34.55%, respectively. Additionally, a weaker QTL on chr3 for the L* value was detected in all three environments. The mean LOD value of the QTL was 4.34, and the mean R^2 was 4.41%, which suggests that this QTL plays a minor role in regulating brown seed coat brightness (Table 3). We continued our analysis of the intervals on chr6 identified by BSA and QTL mapping. Both analysis methods repeatedly identified approximately the same interval. This supports the identification of this interval and its surrounding region as a reliable major-effect QTL controlling brown seed coat traits. The flanking markers chr_16989955 and chr_20193451 spanned a physical distance of 3.2 Mb in the reference genome (chr6: 16.99 Mb~20.19 Mb). Notably, the 1.19 Mb region on chr6 between the markers chr_18323068 and chr_19517928 overlapped with other QTL intervals identified in all environments and is the closest to the LOD peak (Figure 3; Table 3). In summary, by combining BSA and traditional QTL mapping methods, we confirmed the mapping of major-effect QTL regulating the brown coat trait in sesame in the merged region of 18,323,068~19,517,928 bp on chr6, with a physical distance of 1.19 Mb. We designated this QTL *qBSCchr6*.

TABLE 3 QTL information for brown seed coat-related traits detected in the RIL population.

Trait	Environment	chr	Position (cM)	LOD	R^2 (%)	Start (cM)	End (cM)	Physical interval (bp)
L*	SY	3	21.60	4.46	4.31	17.00	23.29	22465300-23218480
	NY		21.60	3.42	4.05	21.60	23.11	22465436-22607606
	ZMD		21.60	5.13	4.87	17.00	23.29	22465300-23218480
	SY	6	93.19	32.84	35.68	93.19	96.87	18323068-19517928
	NY		96.87	12.21	35.79	93.19	101.29	18323068-20193451
	ZMD		93.19	27.76	29.46	89.17	96.87	16989955-19517928
a*	SY	6	93.19	33.97	37.79	93.19	96.87	18323068-19517928
	NY		93.19	35.85	39.62	93.19	96.87	18323068-19517928
	ZMD		93.19	29.25	32.49	89.17	96.87	16989955-19517928
b*	SY	6	93.19	33.74	36.02	89.17	96.87	16989955-19517928
	NY		93.19	34.73	38.60	93.19	96.87	18323068-19517928
	ZMD		93.19	26.06	29.02	89.17	96.87	16989955-19517928

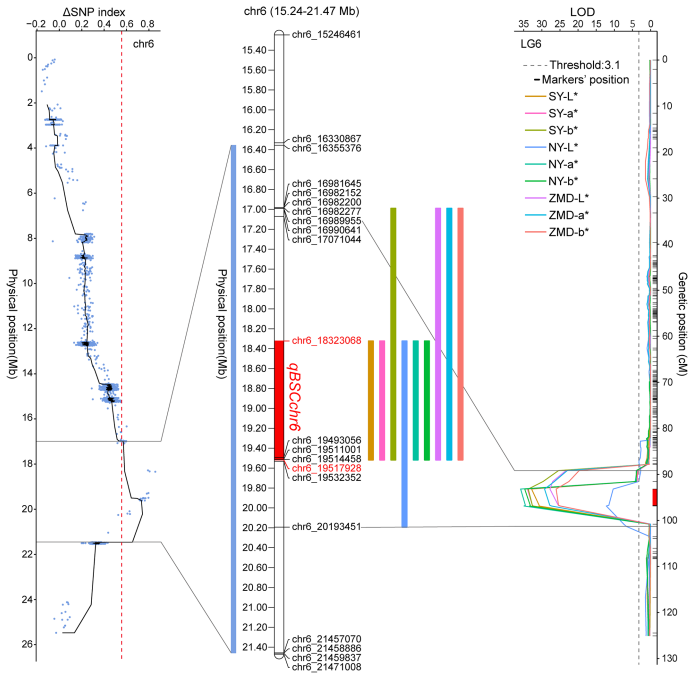


FIGURE 3
Position of *qBSCchr6* on the chr6 physical map. Colored boxes represent the physical distance spanned on chr6 for QTL identified by BSA and QTL mapping.

3.5 Screening of candidate genes and preliminary validation by qRT–PCR

To extract additional information for *qBSCchr6*, we identified a total of 1,720 SNPs/InDels in this interval, among which there were 50 effective SNPs and 16 effective InDels (Supplementary Table 7). In total, there were 118 genes in this candidate region, with intro variants, frameshift variants, disruptive inframe deletions, and missense variants of 45, 8, 4, and 29, respectively (Supplementary Table 8). Ultimately, 42 genes were predicted to show high and moderate variance effects on protein coding (Supplementary Table 9). It was previously reported that seed coat color may be associated with the synthesis of flavonols, anthocyanins, lignin, and melanin (Pourcel et al., 2007; Yu, 2013). We found that 13 of these 118 genes may be associated with brown seed coat color regulation based on their function. Five of these genes showed high or moderate effects on protein coding; SIN_1023218, SIN_1023231, SIN_1023270, and SIN_1023287 were annotated as missense variants, and SIN_1023210 was annotated as a frameshift variant and disruptive in-frame insertion. These variants with high or moderate effects on protein coding may cause the loss of

the original function and thus interrupt the accumulation of pigments in the seed coat (Supplementary Table 10).

Additionally, we observed the phenotypes of the parental characteristics at different days post-anthesis and found that the seed coat color appeared slightly different between the parents starting at 20 DPA, and that some areas of the seeds of YZEHP were colored at 25 DPA and substantially colored at 30 DPA (Figures 4A, B). Next, we performed preliminary qRT–PCR validation of 13 genes with possible functions associated with seed coat color and found that the expression level of SIN_1023239 in YZEHP was significantly up-regulated than YZ8 with 2.5-, 9.4-, 6.0-, and 5.9-fold at 15 DPA, 20 DPA, 25 DPA, and 30 DPA, respectively (Figure 4C). There was no discernible pattern in the expression of the remaining 12 genes in white seeds of YZ8 and brown seeds of YZEHP (Supplementary Figure 8; Supplementary Table 11). Therefore, it was the expression pattern of SIN_1023239 that was consistent with the color accumulation characteristics of the brown seed coat in YZEHP, and thus, it may be crucial for brown seed coloration.

4 Discussion

Seed coat color is a commercially important trait in sesame; seeds with different coat colors show specific characteristics in terms of microelement content, and it aids in the indirect selection of genotypes with high mineral content (Pandey et al., 2017). We performed separate observations and instrumental quantifications of RIL population phenotypes, and used the whole-genome resequencing technique and two computational analysis methods to map QTLs for the sesame brown seed coat trait. *qBSCchr6* was identified as a major-effect QTL that spans a physical interval of 1.19 Mb on chr6. Moreover, based on the effect of gene variants on protein coding and the potential expression pattern of the gene for pigment accumulation during seed coat development, we identified possible candidate genes within this interval.

Laurentin and Benítez (2014) developed four F_2 populations using two white sesame cultivars and one brown sesame cultivar in reciprocal crosses, and phenotypic investigations revealed that all showed consistency with a 3:1 segregation ratio and that brown was dominant to white. This is consistent with our observation that dominant genes controlled the brown seed coat. However, the bimodal continuous distribution of L*, a*, and b* values in the RIL population indicates that a minor-effect genetic locus may also control the brown seed coat trait. Therefore, the use of high-throughput phenotypic data and an increased marker density are both effective ways to improve the efficacy of QTL detection (Li et al., 2010). In addition, the values of L*, a*, and b* obtained in the three environments, showed high heritability. Previous studies have also demonstrated that over 90% of the phenotypic variation in sesame seed coat color is genetically controlled and slightly influenced by environmental factors (Zhang et al., 2013). Moreover, due to indeterminate inflorescence growth, climate, and harvest time, differences in seed maturity at harvest can also cause differences in seed coat color, leading to instability in phenotypic and QTL analyses, as reported based on seed coat color mapping in *Brassica napus* (Yan et al., 2009). Interestingly, in the present study, the mean CV (from high to low) were 27.86%,

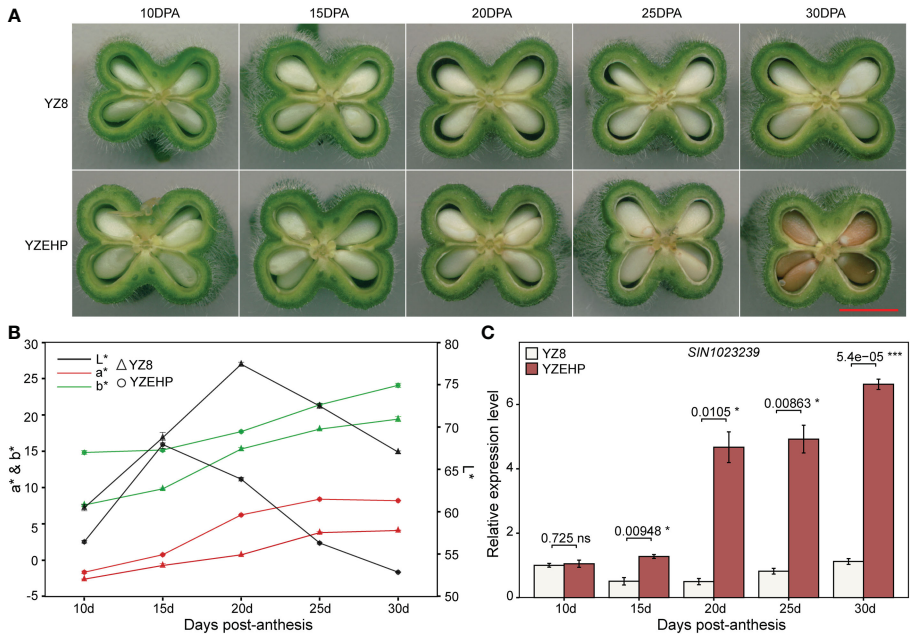


FIGURE 4
Phenotype and gene expression of parental seed coat at different developmental stages. **(A)** Longitudinal sections of capsules, red line segments indicate 0.5 mm. **(B)** Values of L*, a*, and b* for different developmental stages of the parental seed coat. **(C)** Relative expression of SIN_1023239 gene in the two parents. Significant levels of relative gene expression were determined by T-test, with ns, *, and *** representing nonsignificant, and significant at $p<0.05$ and $p<0.001$ levels, respectively.

12.80%, and 6.54% for a*, b*, and L* values, respectively, indicating that a* value had the highest dispersion in the RIL population and best represented the phenotypic characteristics of the brown coat color trait in sesame, while the opposite was true for L* value.

A high-quality genetic map is the basis of QTL mapping for agronomic traits. QTL mapping by whole-genome low coverage sequencing has been successfully applied to chickpea and peanut (Kale et al., 2015; Sun et al., 2022). In these studies, the parental sequencing depths ranged from ~7.9× to 34.58×, the population sequencing depths ranged from 0.72× to 1.4×, and the number of markers used for mapping ranged from ~53,000 to ~210,000. The actual sequencing coverage obtained in whole-genome resequencing averaged 18.61× in the parents and 5.16× in the RIL population, which was considered sufficient for QTL mapping in this study (Supplementary Table 2). Although the number of markers we obtained for mapping was only ~160,000, possibly due to our strict filtering of the marker sequencing depth, this did not affect our ability to construct a reliable and stable genetic map and use it for subsequent QTL mapping. In addition, we found that most of the linkage groups were separated into subgroups due to the uneven distribution of adjacent markers and large gaps (up to ~18 cM), and the calculation of recombination scores was affected by the lack of markers. We further validated collinearity with physical maps (such as LG8, LG10, and LG12) and found that most markers were located in the central region of chromosomes, allowing each chromosome to be split into several contiguous groups, similar to what has been found in wheat and quinoa (Langlands-Perry et al., 2021; Maldonado-Taipe et al., 2022). Importantly, this did not affect our subsequent QTL mapping analysis, which passed several independent tests for quality.

Most previous studies on QTLs regulating sesame seed coat color have included co-mapping for black sesame or segregation of various colors and have not been able to separate the QTLs or mechanisms of interaction mapped to individual seed coat colors. Through 10 successive generations of self-fertilization, we created a population of RILs with stable inheritance and eventually identified a major-effect QTL controlling brown seed coat traits on chr6. Furthermore, we compared *qBSCchr6* with QTLs associated with seed color from previous reports. However, only the results from a genome-wide association study (GWAS) of seed coat color in 366 natural populations included the same physical interval (Cui et al., 2021). In particular, most of the significant SNPs in the GWAS results were mapped to the confidence intervals of *qSCa-4.1*/*qSCb-4.1*/*qSCl-4.1*, *qSCa-8.1*/*qSCb-8.1*/*qSCl-8.1* and *qSCl-8.2* identified by Wang et al. (2016), which further suggests the specificity and accuracy of *qBSCchr6* in controlling brown seed coat color. Other comparable QTLs were not mapped to our confidence interval (Wei et al., 2015; Wang et al., 2016). Some previous studies applied independent genetic maps and genomes, making it difficult to determine the relationships between their results and *qBSCchr6* (Zhang et al., 2013; Du et al., 2019; Li et al., 2021). In the present study, the linkage analysis also revealed a minor-effect QTL for L* color values on chr3 across the three environments, with LOD values between 3.42 and 5.13 and R^2 between 4.05% and 4.87% (Table 3). However, the ΔSNP index in the BSA did not fluctuate within this interval, probably because QTL-seq is not suitable for detecting minor-effect QTLs without the repeated measurement of phenotypes across multiple years (Takagi et al., 2013). Phenotypic data also showed the smallest dispersion of L* values in the three environments, and it is possible that weak changes in brightness do not cause visually detectable differences.

The presentation of seed color in various plants is complex and diverse, involving the main components of flavonols, PAs (concentrated tannins), and some phenolic substances such as lignins and melanins (Yu, 2013). We sampled seeds every 5 days from 10 DPA until we observed significant differences in the seed coat color between the parental plants. From 20 DPA onward, we observed the greatest variation in L* values, with YZEHP seeds being darker than YZ8 seeds, and we eventually noted a clear color difference at 30 DPA. Wang et al. (2020) found that black sesame seeds started to synthesize and accumulate melanin gradually at 8 DPA and that a significant difference in seed coat color appeared at 14 DPA. These results were not exactly the same as ours, and we speculate that this might be due to the different metabolic pathways involved in the accumulation of pigmented substances. A search for candidate genes within the confidence interval of *qBSCchr6* was further performed. Among the 13 screened genes, SIN_1023210 has been annotated as encoding the UDP-glycosyltransferase 87A2 protein associated with catalytic glycosylation (one of the final steps in the production of secondary metabolites) and plays an important role in determining the coloration of flowers, leaves, seeds, and fruits (Le Roy et al., 2016; Foong et al., 2020). SIN_1023231 and SIN_1023270 are annotated as exocyst subcomplex-containing subunit (EXO70) proteins associated with the vesicle-dependent autophagy-related pathway of anthocyanin-containing vesicles from the endoplasmic reticulum into the vesicle lumen (Kulich et al., 2013). SIN_1023248, SIN_1023249, SIN_1023303, and SIN_1023305 all encode peroxidases, which may be related to lignin formation and coloration during fruit ripening (Pourcel et al., 2007; Ring et al., 2013). SIN_1023218 encodes alanine glyoxylate aminotransferase 2, which is involved in the transfer and catalysis of amino acids (Liepman and Olsen, 2003). SIN_1023221 and SIN_1023287 encode 2-oxoglutarate-dependent dioxygenase and beta-glucosidase, respectively, which are essential enzymes in flavonoid and phenylpropanoid biosynthesis (Farrow and Facchini, 2014; Munir et al., 2019). These are all potential regulatory pathways related to seed coat pigment accumulation. Furthermore, SIN_1023237, SIN_1023239, and SIN_1023240 all encode laccase 3 (LAC3), a multicopper glycoprotein that catalyzes and activates the oxidation of diphenol substrates in the presence of molecular oxygen in poplar (Ranocha et al., 1999). However, we found that only SIN_1023239 was significantly up-regulated in YZEHP seeds at different developmental periods compared to its expression in YZ8, and the expression pattern was consistent with the phenotypic trend. In *Arabidopsis*, *TT10* (laccase 15) is involved in the oxidation of concentrated tannins in the seed coat, resulting in brown coat color at harvest, and the other 16 laccase enzymes do not seem to compensate for the loss of activity in the TT10 mutant (Pourcel et al., 2005). In addition, preliminary evidence based on bioinformatics suggests the presence of one or more forms of epigenetic modification in the coding sequences of the eight laccase enzymes including AtLAC3 (Turlapati et al., 2011). In poplar, *LAC3* increased the content of soluble phenols in the seed coat, participated in the oxidation of lignin, and affected the structure and integrity of the cell wall (Ranocha et al., 2002). In maize, *ZmLAC3* is also involved in

the polymerization of phenolic compounds (Caparrós-Ruiz et al., 2006). In addition to flavonoids and anthocyanins, some researchers have surmised that lignins or phenolics affect the seed colors of plants, although the available evidence is not sufficient to support this conclusion (Qu et al., 2013). A recent study by Dossou et al. (2022) focused on the metabolomics of four sesame cultivars and found that the developmental regulation of black, brown, yellow, and white sesame seed coat colors may be different, resulting in different coloration due to variations in the major bioactive phenolic compounds in sesame seeds. Nevertheless, our identification of long fragments of InDels or SNPs may be missed. Further development of markers for fine mapping is needed, and multiomics techniques should be combined to analyze the deposition of sesame seed coat pigments to identify the regulatory mechanisms underlying different color traits.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI–PRJNA934094, SRR23434336 to SRR23434652.

Author contributions

ZW, HM, and BJ directed the project and advised on subsequent studies. HW and CC designed the experiments, performed bioinformatics analysis, completed the construction of the genetic map and computational analysis of genotype frequencies, and completed gene screening and quantitative analysis. YL and YZZ developed the RIL population and performed field experiments. YQZ, XC, and XW performed the DNA extraction. HW together with all authors wrote and finalized the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the China Agriculture Research System (CARS-14–1-01, CARS-14–2-21), the Central Government-Guided Local S&T Development Fund Project of Henan (Z20221343038), the Key Project of Science and Technology of Henan (201300110600), the Key Research and Development Project of Henan (221111520400), and the Key Research and Development Program of Shaanxi (2022NY-073).

Acknowledgments

We would like to thank Dr. Tianxiang Liu and Dr. Hongqi Wu for helpful comments on this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1131975/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Frequency distribution of L*, a*, and b* values in the three environments of the RIL population.

SUPPLEMENTARY FIGURE 2
Statistical information on individual sequencing data of parents and RILs. **(A)** Clean data size distribution. **(B)** Information on the mapped ratio.

SUPPLEMENTARY FIGURE 3
Statistical information on genetic markers used to construct genetic maps. **(A)** Marker type and quantity statistics. **(B)** Statistical information of valid SNP/InDel genetic markers in each linkage group.

SUPPLEMENTARY FIGURE 4
High-density genetic map of RIL population. Each vertical line represents the position of the marker in the linkage groups.

SUPPLEMENTARY FIGURE 5
Haplotype assessment of recombination breakpoints for each sample of the RIL population.

SUPPLEMENTARY FIGURE 6
Analysis of collinearity between genetic and physical maps of sesame. Horizontal coordinates indicate the genetic distance of each linkage group, and vertical coordinates indicate the physical length of each chromosome, and marker collinearity in genomic and genetic maps is represented in the form of scatter.

SUPPLEMENTARY FIGURE 7
Heatmap of pairwise recombination and LOD scores based on 7,817 markers. Estimated recombination scores between markers are shown above the diagonal line, and LOD scores are shown below the diagonal line. Red indicates closely linked markers (high LOD scores and low recombination scores) and blue indicates non-linked markers (low LOD scores and high recombination scores).

SUPPLEMENTARY FIGURE 8
Relative expression levels of candidate genes that were inconsistent with the pattern of phenotypic variation among parents at different developmental stages of the seed coat. Significant levels of relative gene expression differences between parents at each period of seed development were tested by T-test, with ns, *, **, and *** representing nonsignificant, significant at the $p<0.05$, $p<0.01$, and $p<0.001$ level, respectively.

References

Abe, A., Kosugi, S., Yoshida, K., Natsume, S., Takagi, H., Kanzaki, H., et al. (2012). Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat. Biotechnol.* 30 (2), 174–178. doi: 10.1038/nbt.2095

Aruldass, C. A., Venil, C. K., Zakaria, Z. A., and Ahmad, W. A. (2014). Brown sugar as a low-cost medium for the production of prodigiosin by locally isolated serratia marcescens UTM1. *Int. Biodeter. Biodegr.* 95, 19–24. doi: 10.1016/j.ibiod.2014.04.006

Bedigian, D. (2003). Evolution of sesame revisited: domestication, diversity and prospects. *Genet. Resour. Crop Evol.* 50 (7), 779–787. doi: 10.1023/A:1025029903549

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics.* 19 (7), 889–890. doi: 10.1093/bioinformatics/btg112

Caparrós-Ruiz, D., Fornalé, S., Civardi, L., Puigdomènech, P., and Rigau, J. (2006). Isolation and characterisation of a family of laccases in maize. *Plant Sci.* 171 (2), 217–225. doi: 10.1016/j.plantsci.2006.03.007

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly.* 6 (2), 80–92. doi: 10.4161/fly.19695

Cui, C., Liu, Y., Liu, Y., Cui, X., Sun, Z., Du, Z., et al. (2021). Genome-wide association study of seed coat color in sesame (*Sesamum indicum* l.). *PLoS One* 16 (5), e0251526. doi: 10.1371/journal.pone.0251526

Das, D., Datta, A. K., Kumbhakar, D. V., Ghosh, B., and Pramanik, A. (2018). Cytogenetical study of intervarietal hybrids of sesame (*Sesamum indicum* l., pedaliaceae) raised by open pollination. *Cytologia.* 83 (2), 159–163. doi: 10.1508/cytologia.83.159

Debeaujon, I., Peeters, A. J., Léon-Kloosterziel, K. M., and Koornneef, M. (2001). The *TRANSPARENT TESTA12* gene of *Arabidopsis* encodes a multidrug secondary transporter-like protein required for flavonoid sequestration in vacuoles of the seed coat endothelium. *Plant Cell.* 13 (4), 853–871. doi: 10.1105/tpc.13.4.853

Dong, M., Tian, L., Li, J., Jia, J., Dong, Y., Tu, Y., et al. (2022). Improving physicochemical properties of edible wheat gluten protein films with proteins, polysaccharides and organic acid. *LWT-Food Sci. Technol.* 154, 112868. doi: 10.1016/j.lwt.2021.112868

Dossou, S. S. K., Xu, F., You, J., Zhou, R., Li, D., and Wang, L. (2022). Widely targeted metabolome profiling of different colored sesame (*Sesamum indicum* l.) seeds provides new insight into their antioxidant activities. *Food Res. Int.* 151, 110850. doi: 10.1016/j.foodres.2021.110850

Du, H., Zhang, H., Wei, L., Li, C., Duan, Y., and Wang, H. (2019). A high-density genetic map constructed using specific length amplified fragment (SLAF) sequencing and QTL mapping of seed-related traits in sesame (*Sesamum indicum* l.). *BMC Plant Biol.* 19 (1), 588. doi: 10.1186/s12870-019-2172-5

Farrow, S. C., and Facchini, P. J. (2014). Functional diversity of 2-oxoglutarate/Fe (II)-dependent dioxygenases in plant metabolism. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00524

Foong, L. C., Chai, J. Y., Ho, A. S. H., Yeo, B. P. H., Lim, Y. M., and Tam, S. M. (2020). Comparative transcriptome analysis to identify candidate genes involved in 2-methoxy-1,4-naphthoquinone (MNQ) biosynthesis in *Impatiens balsamina* l. *Sci. Rep.* 10 (1), 16123. doi: 10.1038/s41598-020-72997-2

Fuller, D. Q. (2003). Further evidence on the prehistory of sesame. *Asian Agrihist.* 7 (2), 127–137.

Gillman, J. D., Tetlow, A., Lee, J. D., Shannon, J. G., and Bilyeu, K. (2011). Loss-of-function mutations affecting a specific glycine max R2R3 MYB transcription factor result in brown hilum and brown seed coats. *BMC Plant Biol.* 11 (1), 155. doi: 10.1186/1471-2229-11-155

Gonzalez, A., Brown, M., Hatlestad, G., Akhavan, N., Smith, T., Hembd, A., et al. (2016). TTG2 controls the developmental regulation of seed coat tannins in *Arabidopsis* by regulating vacuolar transport steps in the proanthocyanidin pathway. *Dev. Biol.* 419 (1), 54–63. doi: 10.1016/j.ydbio.2016.03.031

Guo, B., Wang, D., Guo, Z., and Beavis, W. D. (2013). Family-based association mapping in crop species. *Theor. Appl. Genet.* 126 (6), 1419–1430. doi: 10.1007/s00122-013-2100-2

Haughn, G., and Chaudhury, A. (2005). Genetic analysis of seed coat development in *Arabidopsis*. *Trends Plant Sci.* 10 (10), 472–477. doi: 10.1016/j.tplants.2005.08.005

Hill, J. T., Demarest, B. L., Bisgrove, B. W., Gorsi, B., Su, Y. C., and Yost, H. J. (2013). MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. *Genome Res.* 23 (4), 687–697. doi: 10.1101/gr.146936.112

Hong, M., Hu, K., Tian, T., Li, X., Chen, L., Zhang, Y., et al. (2017). Transcriptomic analysis of seed coats in yellow-seeded *Brassica napus* reveals novel genes that influence proanthocyanidin biosynthesis. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01674

Hwang, L. S. (2005). Sesame oil. bailey's ind. *Oil Fat Prod.* 2, 547–552. doi: 10.1002/047167849X.bio031

Jia, J., Ji, R., Li, Z., Yu, Y., Nakano, M., Long, Y., et al. (2020). Soybean DICER-LIKE2 regulates seed coat color *via* production of primary 22-nucleotide small interfering RNAs from long inverted repeats. *Plant Cell.* 32 (12), 3662–3673. doi: 10.1105/tpc.20.00562

Kale, S. M., Jaganathan, D., Ruperao, P., Chen, C., Punna, R., Kudapa, H., et al. (2015). Prioritization of candidate genes in "QTL-hotspot" region for drought tolerance in chickpea (*Cicer arietinum* l.). *Sci. Rep.* 5 (1), 15296. doi: 10.1038/srep15296

Kermani, S. G., Saeidi, G., Sabzalian, M. R., and Gianinetti, A. (2019). Drought stress influenced sesamin and sesamolin content and polyphenolic components in sesame (*Sesamum indicum* l.) populations with contrasting seed coat colors. *Food Chem.* 289, 360–368. doi: 10.1016/j.foodchem.2019.03.004

Kulich, I., Pecenkova, T., Sekeres, J., Smetana, O., Fendrych, M., Foissner, I., et al. (2013). *Arabidopsis* exocyst subcomplex containing subunit *EXO70B1* is involved in autophagy-related transport to the vacuole. *Traffic.* 14 (11), 1155–1165. doi: 10.1111/tra.12101

Langlands-Perry, C., Cuenin, M., Bergez, C., Krima, S. B., Gélisse, S., Sourdille, P., et al. (2021). Resistance of the wheat cultivar 'Renan' to septoria leaf blotch explained by a combination of strain specific and strain non-specific QTL mapped on an ultra-dense genetic map. *Genes.* 13 (1), 100. doi: 10.3390/genes13010100

Laurentin, H., and Benítez, T. (2014). Inheritance of seed coat color in sesame. *Pesqui. Agropecu. Bras.* 49, 290–295. doi: 10.1590/S0100-204X2014000400007

Le Roy, J., Huss, B., Creach, A., Hawkins, S., and Neutelings, G. (2016). Glycosylation is a major regulator of phenylpropanoid availability and biological activity in plants. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00735

Li, C., Duan, Y., Miao, H., Ju, M., Wei, L., and Zhang, H. (2021). Identification of candidate genes regulating the seed coat color trait in sesame (*Sesamum indicum* l.) using an integrated approach of QTL mapping and transcriptome analysis. *Front. Genet.* 12. doi: 10.3389/fgene.2021.700469

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics.* 25 (14), 1754–1760. doi: 10.1093/bioinformatics/btp324

Liepman, A. H., and Olsen, L. J. (2003). Alanine aminotransferase homologs catalyze the glutamate: glyoxylate aminotransferase reaction in peroxisomes of *Arabidopsis*. *Plant Physiol.* 131 (1), 215–227. doi: 10.1104/pp.011460

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25 (16), 2078–2079. doi: 10.1093/bioinformatics/btp352

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods.* 25 (4), 402–408. doi: 10.1006/meth.2001.1262

Li, H., Zhang, L., and Wang, J. (2010). Analysis and answers to frequently asked questions in quantitative trait locus mapping. *Acta Agron. Sin.* 36 (6), 918–931. doi: 10.3724/SP.J.1006.2010.00918

Maldonado-Taipe, N., Barbier, F., Schmid, K., Jung, C., and Emrani, N. (2022). High-density mapping of quantitative trait loci controlling agronomically important traits in quinoa (*Chenopodium quinoa* willd.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.916067

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20 (9), 1297–1303. doi: 10.1101/gr.107524.110

Mei, H., Liu, Y., Du, Z., Wu, K., Cui, C., Jiang, X., et al. (2017). High-density genetic map construction and gene mapping of basal branching habit and flowers per leaf axil in sesame. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00636

Meng, L., Li, H. H., Zhang, L. Y., and Wang, J. K. (2015). QTL IciMapping: Integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3 (3), 269–283. doi: 10.1016/j.cj.2015.01.001

Munir, N., Cheng, C., Xia, C., Xu, X., Nawaz, M. A., Iftikhar, J., et al. (2019). RNA-Seq analysis reveals an essential role of tyrosine metabolism pathway in response to root-rot infection in *Gerbera hybrida*. *PLoS One* 14 (10), e0223519. doi: 10.1371/journal.pone.0223519

Nordnorg, M., and Welgel, D. (2008). Next-generation genetics in plants. *Nature.* 456, 720–723. doi: 10.1038/nature07629

Oren, E., Tzuri, G., Dafna, A., Rees, E. R., Song, B., Freilich, S., et al. (2022). QTL mapping and genomic analyses of earliness and fruit ripening traits in a melon recombinant inbred lines population supported by *de novo* assembly of their parental genomes. *Hortic. Res. -England.* 9. doi: 10.1093/hr/uhab081

Ouellette, L. A., Reid, R. W., Blanchard, S. G., and Brouwer, C. R. (2018). LinkageMapView–rendering high-resolution linkage and QTL maps. *Bioinformatics.* 34 (2), 306–307. doi: 10.1093/bioinformatics/btx576

Pandey, S. K., Das, A., and Dasgupta, T. (2013). Genetics of seed coat color in sesame (*Sesamum indicum* l.). *Afr. J. Biotechnol.* 12 (42), 6061–6067. doi: 10.5897/AJB2013.13055

Pandey, S. K., Majumder, E., and Dasgupta, T. (2017). Genotypic variation of microelements concentration in sesame (*Sesamum indicum* l.) mini core collection. *Agric. Res.* 6, 114–121. doi: 10.1007/s40003-017-0252-z

Pourcel, L., Routaboul, J. M., Cheynier, V., Lepiniec, L., and Debeaujon, I. (2007). Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* 12 (1), 29–36. doi: 10.1016/j.tplants.2006.11.006

Pourcel, L., Routaboul, J. M., Kerhoas, L., Caboche, M., Lepiniec, L., and Debeaujon, I. (2005). *TRANSPARENT TESTA10* encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in *Arabidopsis* seed coat. *Plant Cell.* 17 (11), 2966–2980. doi: 10.1105/tpc.105.035154

Prasad, R., and Gangopadhyay, G. (2011). Phenomic analyses of Indian and exotic accessions of sesame (*Sesamum indicum* l.). *J. Plant Breed. Crop Sci.* 3 (13), 336–352. doi: 10.5897/JPBCS11.049

Qu, C., Fu, F., Lu, K., Zhang, K., Wang, R., Xu, X., et al. (2013). Differential accumulation of phenolic compounds and expression of related genes in black- and yellow-seeded *Brassica napus*. *J. Exp. Bot.* 64 (10), 2885–2898. doi: 10.1093/jxb/ert148

Ranocha, P., Chabannes, M., Chamayou, S., Danoun, S., Jauneau, A., Boudet, A. M., et al. (2002). Laccase down-regulation causes alterations in phenolic metabolism and cell wall structure in poplar. *Plant Physiol.* 129 (1), 145–155. doi: 10.1104/pp.010988

Ranocha, P., McDougall, G., Hawkins, S., Sterjiades, R., Borderies, G., Stewart, D., et al. (1999). Biochemical characterization, molecular cloning and expression of laccases–a divergent gene family–in poplar. *Eur. J. Biochem.* 259 (1-2), 485–495. doi: 10.1046/j.1432-1327.1999.00061.x

Ren, Y., He, Q., Ma, X., and Zhang, L. (2017). Characteristics of color development in seeds of brown- and yellow-seeded heading chinese cabbage and molecular analysis of *Brsc*, the candidate gene controlling seed coat color. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01410

Ring, L., Yeh, S. Y., Hucherig, S., Hoffmann, T., Blanco-Portales, R., Fouche, M., et al. (2013). Metabolic interaction between anthocyanin and lignin biosynthesis is associated with peroxidase *FaPRX27* in strawberry fruit. *Plant Physiol.* 163 (1), 43–60. doi: 10.1104/pp.113.222778

Shahidi, F., Liyana-Pathirana, C. M., and Wall, D. S. (2006). Antioxidant activity of white and black sesame seeds and their hull fractions. *Food Chem.* 99 (3), 478–483. doi: 10.1016/j.foodchem.2005.08.009

Sun, Z. Q., Qi, F. Y., Liu, H., Qin, L., Xu, J., Shi, L., et al. (2022). QTL mapping of quality traits in peanut using whole-genome resequencing. *Crop J.* 10 (1), 177–184. doi: 10.1016/j.cj.2021.04.008

Takagi, H., Abe, A., Yoshida, K., Kosugi, S., Natsume, S., Mitsuoka, C., et al. (2013). QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* 74 (1), 174–183. doi: 10.1111/tpj.12105

Tao, J., Li, S., Wang, Q., Yuan, Y., Ma, J., Xu, M., et al. (2022). Construction of a high-density genetic map based on specific-locus amplified fragment sequencing and identification of loci controlling anthocyanin pigmentation in yunnan red radish. *Hortic. Res. -England.* 9. doi: 10.1093/hr/uhab031

Taylor, J., and Butler, D. (2017). R package ASMap: Efficient genetic linkage map construction and diagnosis. *J. Stat. Software* 79 (6), 1–29. doi: 10.18637/jss.v079.i06

Teshima, T. (1931). Inheritance of the color of seed coat in sesame. *Jpn. J. Crop Sci.* 3 (3), 232–235. doi: 10.1626/jcs.3.232

Turlapati, P. V., Kim, K.-W., Davin, L. B., and Lewis, N. G. (2011). The *laccase* multigene family in *Arabidopsis thaliana*: towards addressing the mystery of their gene function (s). *Planta.* 233 (3), 439–470. doi: 10.1007/s00425-010-1298-3

Wang, L., Dossou, S. S. K., Wei, X., Zhang, Y., Li, D., Yu, J., et al. (2020). Transcriptome dynamics during black and white sesame (*Sesamum indicum* l.) seed development and identification of candidate genes associated with black pigmentation. *Genes.* 11 (12), 1399. doi: 10.3390/genes11121399

Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., et al. (2016). Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. *BMC Genomics* 17 (1), 31. doi: 10.1186/s12864-015-2316-4

Wei, X., Liu, K., Zhang, Y., Feng, Q., Wang, L., Zhao, Y., et al. (2015). Genetic discovery for oil production and quality in sesame. *Nat. Commun.* 6 (1), 8609. doi: 10.1038/ncomms9609

Wei, X., Zhu, X., Yu, J., Wang, L., Zhang, Y., Li, D., et al. (2016). Identification of sesame genomic variations from genome comparison of landrace and variety. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01169

Wickham, H. (2016). "Data analysis," in *ggplot2* (Cham: Springer). doi: 10.1007/978-3-319-24277-4_9

Wu, Y., Bhat, P. R., Close, T. J., and Lonardi, S. (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* 4 (10), e1000212. doi: 10.1371/journal.pgen.1000212

Xu, J., Chen, S., and Hu, Q. (2005). Antioxidant activity of brown pigment and extracts from black sesame seed (*Sesamum indicum* l.). *Food Chem.* 91 (1), 79–83. doi: 10.1016/j.foodchem.2004.05.051

Yan, X. Y., Li, J. N., Fu, F. Y., Jin, M. Y., Chen, L., and Liu, L. Z. (2009). Co-Location of seed oil content, seed hull content and seed coat color QTL in three different environments in *Brassica napus* l. *Euphytica.* 170 (3), 355–364. doi: 10.1007/s10681-009-0006-5

Yu, C. Y. (2013). Molecular mechanism of manipulating seed coat coloration in oilseed *Brassica* species. *J. Appl. Genet.* 54 (2), 135–145. doi: 10.1007/s13353-012-0132-y

Zhang, J., Lu, Y., Yuan, Y., Zhang, X., Geng, J., Chen, Y., et al. (2009). Map-based cloning and characterization of a gene controlling hairiness and seed coat color traits in *Brassica rapa*. *Plant Mol. Biol.* 69 (5), 553–563. doi: 10.1007/s11103-008-9437-y

Zhang, H., Miao, H., Wei, L., Li, C., Zhao, R., and Wang, C. (2013). Genetic analysis and QTL mapping of seed coat color in sesame (*Sesamum indicum* l.). *PLoS One* 8 (5), e63898. doi: 10.1371/journal.pone.0063898

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 12 April 2023
DOI 10.3389/fpls.2023.1174985

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Jianfeng Weng,
Institute of Crop Sciences (CAAS), China
Muhammad Irfan Siddique,
North Carolina State University,
United States

*CORRESPONDENCE
Hui Liao
nkliaohui@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science

RECEIVED 27 February 2023
ACCEPTED 28 March 2023
PUBLISHED 12 April 2023

CITATION
Zhang X, Wang M, Guan H, Wen H, Zhang C, Dai C, Wang J, Pan B, Li J and Liao H (2023) Genetic dissection of QTLs for oil content in four maize DH populations.
Front. Plant Sci. 14:1174985.
doi: 10.3389/fpls.2023.1174985

Genetic dissection of QTLs for oil content in four maize DH populations

Xiaolei Zhang[1†], Min Wang[2†], Haitao Guan[1], Hongtao Wen[1], Changzheng Zhang[3], Changjun Dai[1], Jing Wang[1], Bo Pan[1], Jialei Li[4] and Hui Liao[1*]

[1]Quality and Safety Institute of Agricultural Products, Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang, China, [2]National Maize Improvement Center of China, College of Agronomy and Biotechnology, China Agricultural University, Beijing, China, [3]Maize Yufeng Biotechnology LLC, Beijing, China, [4]Food Processing Institute, Heilongjiang Academy of Agricultural Sciences, Harbin, Heilongjiang, China

Oil is one of the main components in maize kernels. Increasing the total oil content (TOC) is favorable to optimize feeding requirement by improving maize quality. To better understand the genetic basis of TOC, quantitative trait loci (QTL) in four double haploid (DH) populations were explored. TOC exhibited continuously and approximately normal distribution in the four populations. The moderate to high broad-sense heritability (67.00-86.60%) indicated that the majority of TOC variations are controlled by genetic factors. A total of 16 QTLs were identified across all chromosomes in a range of 3.49-30.84% in term of phenotypic variation explained. Among them, six QTLs were identified as the major QTLs that explained phenotypic variation larger than 10%. Especially, *qOC-1-3* and *qOC-2-3* on chromosome 9 were recognized as the largest effect QTLs with 30.84% and 21.74% of phenotypic variance, respectively. Seventeen well-known genes involved in fatty acid metabolic pathway located within QTL intervals. These QTLs will enhance our understanding of the genetic basis of TOC in maize and offer prospective routes to clone candidate genes regulating TOC for breeding program to cultivate maize varieties with the better grain quality.

KEYWORDS

Maize, DH, kernel, oil, QTL

1 Introduction

The modern maize (*Zea mays* L.) kernels are composed of approximately 72% starch, 10% protein, 4% oil, and 14% other constituents (Laurie et al., 2004; Ranum et al., 2014). Oil predominantly accumulates in the embryo and is stored in the form of triacylglycerols, which is composed of roughly 59% polyunsaturated, 24% monounsaturated and 13% saturated fatty acid (Dupont et al., 1990; Lambert, 2001). The proper ratio of unsaturated to saturated fatty acids in maize oil is considered as a character of high-quality oil for human

health (Han et al., 1987; Benitez et al., 1999; Lambert et al., 2004). In addition, the high energy and proportion of polyunsaturated fatty acids is highly valued for animal feed, industrial applications and an alternative to fossil fuels (Hou et al., 2022). Thus, the ability to improve oil quantity and quality has been a key target for plant breeding and biotechnology-assisted improvement (Yang et al., 2012; Li et al., 2013).

High-oil maize hybrids (oil concentration > 6%) are considered as an important crop with valued nutrient (Wei et al., 2009). A series of genetic resources have been generated by long-term artificial selection of high-oil maize populations (Fang et al., 2021). The oil concentration of initial open-pollinated variety Illinois High Oil (IHO) reached about 20% after 100 generations of selection (Dudley and Lambert, 2004). A normal maize synthetic Zhongzong No. 2, which was synthesized with 12 inbred lines of Lancast heterotic group, was used to produce the Beijing High Oil (BHO) with oil concentration increased from 4.71 to 15.55% after 18 selection cycles (Song and Chen, 2004). The inbred line By804 was derived from the high-oil population 'Beinongda' and its oil concentration reached 11.22% (Zhang et al., 2008).

As the unique and precious resources, these high oil materials provide an opportunity to understand the genetic architecture of oil and fatty acid biosynthesis, which in turn increase the efficiency of selection to improve oil concentration and quality (Wassom et al., 2008a; Wassom et al., 2008b; Yang et al., 2010; Li et al., 2020). Combined with map-based cloning, QTL mapping is the most powerful and efficient strategy to identify the genomic region that controls complex quantitative traits in plants (Goldman et al., 1994; Lima et al., 2006; Messmer et al., 2009). The total oil content (TOC) is a quantitative trait, and many quantitative trait loci (QTL) have been demonstrated to control the seed oil accumulation in a randomly mated $F_{2:3}$ population IHO × ILO (Alrefai et al., 1995; Berke and Rocheford, 1995; Laurie et al., 2004; Clark et al., 2006; Dudley, 2008). These studies revealed that TOC was controlled by numerous genes with individually small effects and mainly additive gene action (Yang et al., 2010). In addition, using a recombinant inbred line (RIL) population derived from B73 × By804, a relatively small number of QTL were detected and accounted for a large percentage of the total phenotypic variation (Song and Chen, 2004; Zhang et al., 2008; Yang et al., 2010; Pan et al., 2012; Yang et al., 2012). These studies also indicated that epistasis is a key factor affecting the genetic basis of oil content in maize kernel (Wassom et al., 2008b; Yang et al., 2010). Similar results were also obtained in two publicly available maize genetic resources, NAM (the nested association mapping population) and AMP508 (association mapping population) based on high-resolution and high power QTL analysis (Lambert et al., 2004; Cook et al., 2012). A high-oil QTL (*qHO6*) on chromosome 6 has been cloned and the candidate gene encodes an acyl-CoA: diacylglycerol acyltransferase (DGAT1-2), which catalyzes the final step of oil synthesis (Zheng et al., 2008). The major QTL QTL-*Pal9* explaining 42% of the phenotypic variation in palmitic acid content was identified on maize chromosome 9 in a bi-parental segregating population and the candidate gene *Zmfatb* encodes acyl-ACP thioesterase (Li et al., 2011).

Distinct mapping populations were featured with advantages and limitations, which results in significant impacts on QTL outputs (Odell et al., 2022). DH segregating populations have been commonly used in QTL analysis for several specific advantages (Chaikam et al., 2019). Complete homozygosity of DH lines allows accurate phenotyping over multiple locations and years compared to families in early selfing generations (Foiada et al., 2015; Yan et al., 2017). In this study, we utilized four DH populations derived from the practical breeding program to further dissect the genetic basis and QTLs controlling the phenotypic variation of TOC in maize kernels. Our intention was to describe the genetic architecture of oil variation in extensive scale and provide the prospective targets to identify candidate genes for increasing oil concentration in commercial maize germplasms.

2 Materials and methods

2.1 Plant materials and field experiments

Four DH populations (TOC1, TOC2, TOC3 and TOC4) were constructed as previously method described (Chaikam et al., 2019; Du et al., 2020). The eight inbred parental lines exhibiting the variation in TOC (Table 1) were belonged to Maize Yufeng Biotechnology LLC (Beijing, China) and selected as elite inbred lines used for optimizing grain nutritional quality breeding program. Parents of TOC1 and TOC2 belong to maize Lancaste germplasm, and parents of TOC3 and TOC4 belong to Reid Yellow Dent germplasm. The populations (TOC1, TOC2, TOC3 and TOC4) including 123, 129, 281 and 160 lines, respectively (Table 1). Each population with its parents were planted in 2021 at Liaoning province, China (LN, 40°`82′N, 123°56′E) with three replication blocks. All lines were planted in a single row plot with the length of 150 cm and 60 cm using a complete randomized block design under natural field conditions. All plants were self-pollinated and kernels from middle part of three well-grown ears were harvested and dried for oil measurement. We declare that all the collections of plant and seed specimens related to this study were performed in accordance with the relevant guidelines and regulations by Ministry of Agriculture (MOA) of the People's Republic of China.

2.2 Evaluation of oil content and statistical analysis of phenotypic data

Near infrared reflectance (NIR) spectrometer (DA 7250, Perten Instruments Inc., Sweden) was used to measure TOC in maize kernels as previously described with a few modifications (Chen and Hu, 2017). The reflectance spectra were collected in a range of 400 to 2500 nm with 10-nm intervals in the NIR region. A minimum of 50 kernels per sample was scanned three times and the average was taken as final phenotypic value.

All statistical analyses were performed by using R Version 4.0.1 (www.R-project.org) as previously described (Zhang et al., 2021; Zhang et al., 2022). The R 'AOV' function was used to estimate the variances of TOC. The model for the variance analysis was as following: $y = \mu + \alpha_g + \beta_e + \epsilon$, where α_g is the effect of the g^{th} line, β_e

TABLE 1 Phenotypic performance, variance, and broad-sense heritability of TOC in the four DH populations.

Trait [a]	Populations							
	TOC1		TOC2		TOC3		TOC4	
Parents								
means ± SD (%)	KB717001	4.14 ± 0.17	KB717001	4.14 ± 0.17	AJ519002	4.30 ± 0.10	AJ519004	4.43 ± 0.02
	KB519009	3.50 ± 0.15	KB719010	3.16 ± 0.05	AJ519001	4.90 ± 0.09	AJ519006	4.95 ± 0.09
***p* value [b]**	0.008**		0.006**		0.002**		0.007**	
DHs								
Size	123		129		281		160	
means ± SD (%)	4.57 ± 0.41		4.42 ± 0.40		4.50 ± 0.42		5.02 ± 0.41	
Range (%)	3.64 - 5.58		3.59 - 5.48		3.10 - 5.42		4.06 - 6.13	
σ_g^2 [c]	0.205		0.183		0.186		0.168	
σ_e^2 [d]	0.027		0.059		0.023		0.009	
σ_ϵ^2 [e]	0.126		0.085		0.274		0.197	
h^2 (%) [f]	83.00%		86.60%		67.00%		71.80%	

[a]TOC;
[b]P value based on a t-test evaluating two parental lines;
[c]genetic variance;
[d]environmental variance;
[e]residual variance;
[f]broad-sense heritability (h^2);
** $p \leq 0.01$.

is the effect of the e^{th} environment, and ϵ is the error. The effects in the model were defined by random. The broad-sense heritability (h^2) analyzed in the populations was calculated according to Knapp et al., 1985. The formula was $h^2 = \sigma_g^2/(\sigma_g^2 + \sigma_\epsilon^2/e)$, where σ_g^2 is the genetic variance, σ_ε^2 is the residual error, and e is the number of environments. The best linear unbiased predictor (BLUP) value of each line was calculated as: $y_{ij} = \mu + e_i + f_j + \epsilon_{ij}$, where y_{ij} is the phenotypic value of individual j in environment i, μ is the grand mean, e_i is the effect of different environments, f_j is the genetic effect, and ϵ_{ij} is the random error. The grand mean was fitted as a fixed effect, and genotype and environment were considered random effects (Wang et al., 2015). All of these variances were estimated using the 'LME4' R package. The BLUP values were used for phenotypic description statistics and QTL analysis.

2.3 Genotyping and constructing genetic linkage map

The four DH populations with their parents were genotyped using the GenoBaits Maize 1K marker panel (Mol Breeding Biotechnology Co., Ltd., Shijiazhuang, China). A total of 4,589 SNP markers were identified on the basis of genotyping by target sequencing platform (Guo et al., 2019). The minor allele frequency (MAF) and missing rate were estimated in each population and the SNPs with MAF < 0.1 or missing rate > 0.6 were filtered out. After quality control, the polymorphic SNPs between two parental lines were used to construct the genetic linkage maps using the R/qtl package functions est.rf and est.map (Broman et al., 2003) with the kosambi mapping method.

2.4 QTL mapping

Composite interval mapping (CIM) method followed by multiple QTL mapping analysis was performed using Windows QTL Cartographer 2.5 and R language (Wang et al., 2010a). The whole genome was scanned at every 1.0 cM interval with a window size of 10 cM. A forward and backward stepwise regression with five controlling markers was conducted to control background from flanking markers. The empirical logarithm of the odds (LOD) threshold was calculated using 1,000 permutations at a significance level of $p = 0.05$ (Churchill and Doerge, 1994). These threshold LOD values were in a range of 2.76 to 3.06 in four DH populations. QTLs with LOD value greater than the threshold were considered for further analysis. With the 1.5-LOD support interval method, the confidence interval for each QTL position was estimated (Lander and Botstein, 1989). The additive × additive epistatic interactions was performed by "IM-EPI" method in IciMapping Version 4.2.

2.5 Gene annotation

QTLs were delimited to a single peak bin interval based on bin map. The protein-coding genes within intervals were listed

according to MaizeGDB database (V2). Each of the corresponding gene were annotated by performing BLASTP searches at the NCBI (blast.ncbi.nlm.nih.gov/Blast.cgi).

3 Results

3.1 Phenotypic variation and heritability of TOC in maize kernel

Four DH populations, TOC1-TOC4 were developed from eight inbred lines (TOC with a range of 3.16-4.95%). Each population contained 123-281 lines, respectively (Table 1). Within each DH population, TOC exhibited a continuously and approximately normal distribution, which is the typical characteristic of quantitative trait (Figure 1 and Table 1). Analysis of variance (ANOVA) revealed that the genotype variance was greater than environmental variance in all populations (Table 1), indicating that phenotypic variations were mainly controlled by genetic factors. Broad-sense heritability estimates were calculated and showed high for TOC1 and TOC2 populations (83.00-86.60%), and moderate for TOC3 and TOC4 populations (67.00-71.80%) (Table 1). The moderate to high heritability indicated that most of TOC variations in these DH populations were genetically controlled and suitable for further QTL mapping.

3.2 Genotyping and genetic linkage map

A GenoBaits Maize 1K SNP marker panel was used for genotyping all DH lines in the four populations. After quality control, a total of 1,217, 575, 1,022 and 1,039 polymorphic SNPs were identified for TOC1-TOC4 populations, respectively. These high-fidelity SNPs were used to construct the genetic linkage map with the missing rate in most lines less than 2% (Figure S1). In total, 925.92, 684.23, 860.81 and 836.67 cM genetic distances spanned in four linkage maps (Figure S2), and the average genetic distance between every two adjacent markers was 0.77, 1.21, 0.85, and 0.81 cM in each DH population, respectively (Table S1).

3.3 Identification of QTLs for TOC in four DH populations

A total of 16 QTLs were identified with a LOD threshold of above 3.00 at the 0.05 significance level (Table 2 and Figure 2). Among them, 3, 4, 5 and 4 QTLs were detected in TOC1, TOC2, TOC3 and TOC4, respectively. The average genetic intervals of these QTLs was 82.69 cM in a range of 36.56-125.29 cM. The average physical interval was 102.58 Mb in a range of 11.96-232.42 Mb. The contribution to phenotypic variation for each population ranged from 40.99 (TOC3) to 62.05% (TOC2) with an average of 51.10%. The explained phenotypic variation were less than broad-sense heritability (Tables 1, 2), suggesting that only part of QTLs have been detected in these bi-parent populations.

In TOC1, three QTLs (*qOC-1-1, qOC-1-2 and qOC-1-3*) distributed on chromosome 3, 5 and 9. The QTL, *qOC-1-3*, with the largest effect (30.84% of the phenotypic variation) was located on chromosome 9. The parental KB717001 allele at this locus had an additive effect of 0.24% for increased oil content. The second QTL *qOC-1-2* was located on chromosome 5, and explained 11.64% of phenotypic variance with an additive effect of 0.15%. *qOC-1-1* on chromosome 3 explained 7.50% of the phenotypic variance and considered as a minor QTL. The parent KB717001 allele at all of mapped loci had increasing effects for TOC.

In TOC2, four QTLs (*qOC-2-1*, *qOC-2-2*, *qOC-2-3* and *qOC-2-4*) were identified and accounted for 62.50% of the total phenotypic variance. One major QTL *qOC-2-3* located on chromosome 9 and contributed to 21.74% of the explained phenotypic variance. The second QTL *qOC-2-2* on chromosome 2 explained 13.53% of phenotypic variance with an additive effect of 0.15%. The *qOC-2-1* and *qOC-2-4* explained 5.72% and 7.26% of the phenotypic variance, respectively The parent KB717001 allele increased the TOC for *qOC-2-1*, *qOC-2-2* and *qOC-2-3*, but decreased the TOC for *qOC-2-4*.

In TOC3, a total of five QTLs (*qOC-3-1*, *qOC-3-2*, *qOC-3-3*, *qOC-3-4* and *qOC-3-5*) were detected and explained 40.99% of the total phenotypic variance. *qOC-3-3* on chromosome 4 was the major QTL explaining phenotypic variation of 12.99% with an additive effect of 0.15%. The parent AJ519002 allele at *qOC-3-2*

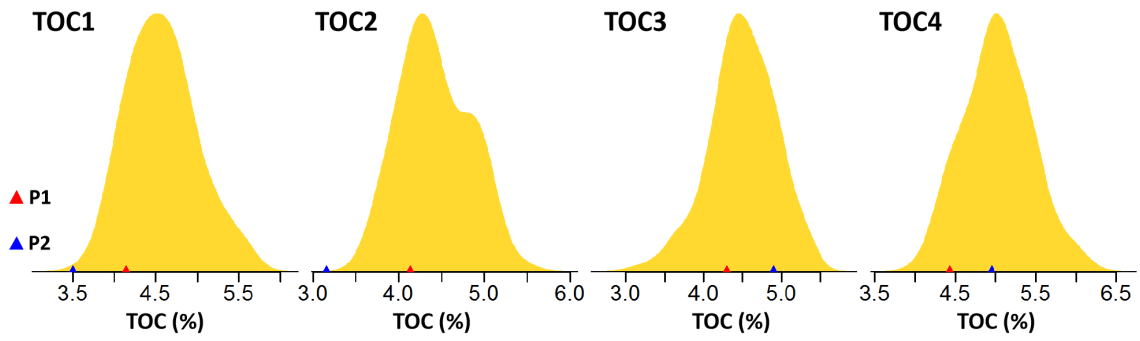


FIGURE 1
Phenotypic variation in TOC in the four DH populations. The *x*-axis showed the TOC and the triangle color indicated the TOC in parents.

10.3389/fpls.2023.1174985

TABLE 2 Individual QTL for TOC in the four DH populations.

Populations	QTL	Chr.[a]	G-Peak (cM)[b]	P-Peak (Mb) _V4[c]	G-Range (cM)[d]	P-Range (Mb) _V4[e]	LOD	PVE %[f]	Add.[g]	Parent[h]+	PVE(%) -ALL[i]
TOC1	*qOC-1-1*	3	48.55	162.54	41.05-52.23	112.37-169.00	4.68	7.50	0.12	KB717001	55.82
	qOC-1-2	5	53.92	193.53	50.84-58.39	191.47-199.14	5.61	11.64	0.15	KB717001	
	qOC-1-3	9	34.02	125.24	31.10-41.66	113.85-143.02	12.24	30.84	0.24	KB717001	
TOC2	*qOC-2-1*	1	19.15	12.72	8.15-29.13	12.72-26.18	4.17	7.26	0.11	KB717001	62.05
	qOC-2-2	2	20.42	30.39	13.01-24.42	11.17-30.39	8.67	13.53	0.15	KB717001	
	qOC-2-3	9	27.23	129.70	25.45-28.83	122.00-130.80	12.80	21.74	0.20	KB717001	
	qOC-2-4	10	23.05	55.99	23.05-24.63	55.99-79.47	4.03	5.72	-0.10	KB719010	
TOC3	*qOC-3-1*	2	42.18	58.26	40.75-42.18	46.13-58.26	3.49	3.49	-0.08	AJ519001	40.99
	qOC-3-2	3	37.82	22.64	29.65-44.47	11.58-149.70	7.37	8.39	0.12	AJ519002	
	qOC-3-3	4	47.22	232.42	43.03-56.33	196.02-241.81	11.71	12.99	-0.15	AJ519001	
	qOC-3-4	5	30.81	43.18	23.31-38.99	15.74-85.58	5.33	5.41	-0.10	AJ519001	
	qOC-3-5	5	65.51	202.27	58.23-79.97	188.27-207.38	7.58	8.26	-0.12	AJ519001	
TOC4	*qOC-4-1*	5	24.43	11.96	15.65-31.84	6.09-20.79	5.12	8.84	-0.13	AJ519006	45.54
	qOC-4-2	6	36.45	131.71	35.42-45.25	129.11-140.52	6.83	13.05	-0.15	AJ519006	
	qOC-4-3	7	54.69	165.51	54.69-54.69	146.02-168.32	3.04	5.07	-0.10	AJ519006	
	qOC-4-4	8	21.13	63.28	19.12-22.76	10.65-65.48	8.82	16.20	-0.18	AJ519006	

[a]Chromosome;
[b]Genetic position in centimorgans (cM) of QTL with the highest LOD;
[c]Physical position of QTL based on the B73 reference sequence (V4);
[d]Genetic position range in centimorgans (cM) of QTL with the highest LOD;
[e]Physical position range of QTL based on the B73 reference sequence (V4);
[f]Percentage of the phenotypic variation explained by the additive effect of QTL;
[g]Additive effect of QTL;
[h]which parental allele increased the TOC;
[i]Percentage of the phenotypic variation explained by the additive effect of all QTL.

increased the TOC, whereas the parent AJ519001 allele at other QTLs increased the TOC.

In TOC4, a total of four QTLs were identified (*qOC-4-1*, *qOC-4-2*, *qOC-4-3* and *qOC-4-4*) and accounted for 45.54% of the total phenotypic variance. *qOC-4-2* on chromosome 6 and *qOC-4-3* on chromosome 8 were two major QTLs explaining the phenotypic variation of 13.05% and 16.20%, respectively. *qOC-4-1* and *qOC-4-3* were two minor QTLs explaining 8.84% and 5.07% phenotypic variation, respectively. The parent AJ519006 allele at all these QTLs increased the TOC.

3.4 Genetic overlap of QTLs in the four DH populations with other populations

Several overlapped QTLs regions were detected across the four populations, including a 37.32 Mb overlap between *qOC-1-1* and *qOC-3-2*, and a 5.05 Mb overlap between *qOC-3-4* and *qOC-4-1* (Figure 3). Moreover, *qOC-1-2* and *qOC-2-3* located within *qOC-3-5* and *qOC-1-3*, respectively (Figure 3).

To investigate whether these newly-identified QTLs shared across different genetic background, we compared their genomic

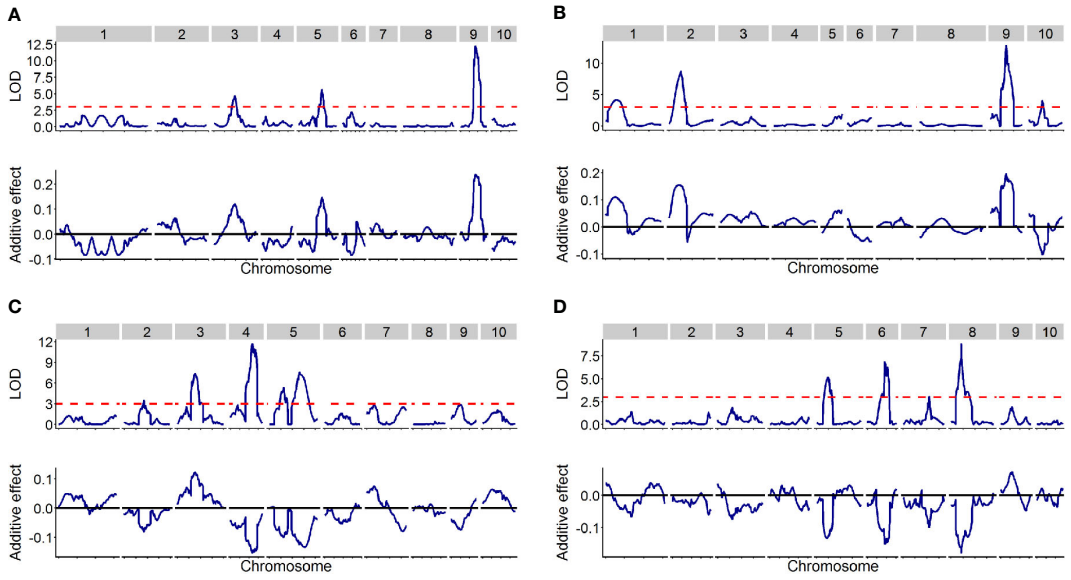


FIGURE 2
The distribution of QTLs across the entire genome in the four DH populations. The upper of each picture displayed LOD score (*y*-axis) against the physical position (*x*-axis) of markers, while the bottom of the picture displayed additive effect (*y*-axis) against the physical position (*x*-axis) of markers. **(A–D)** designated TOC1, TOC2, TOC3 and TOC4, respectively.

locations with QTLs related to oil traits from the other eight previous studies (Mangolin et al., 2004; Wassom et al., 2008a; Wassom et al., 2008b; Wang et al., 2010b; Cook et al., 2012; Yang et al., 2012; Li et al., 2013; Yang et al., 2016; Karn et al., 2017; Fang et al., 2020 and Fang et al., 2021). A total of 56 genomic regions related to oil synthesis and accumulations were identified to be overlapped with QTLs in our four DH populations (Figure 3). These results indicated that although unique and specific QTLs were detected in each population, some genetic loci may have common effects on TOC among different types of populations.

4 Discussion

4.1 QTL mapping precision

The genetic architecture of a quantitative trait consists of a set of parameters that explain the genetic component of trait variation within or among populations (Laurie et al., 2004). These parameters include the number of QTL affecting the trait, their locations in the genome, the frequencies of alternative genotypes segregating at the QTL, the pattern of linkage disequilibria among QTL, and the

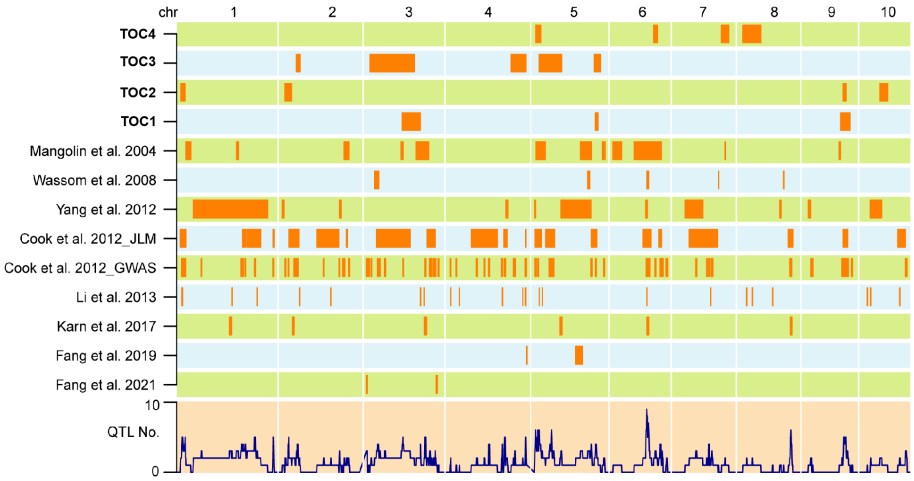


FIGURE 3
Co-localization of TOC QTLs in maize kernels identified in the present and previous studies. The QTLs identified in this study were represented on top. QTLs detected in previous studies were displayed in the form of references. The lower layer showed the number of detected QTLs.

magnitudes of additive, dominance, and epistatic effects (Laurie et al., 2004). Different types of populations used in QTL mapping tend to vary with two main characteristics: (1) their ability to capture genetic diversity, and (2) their power to detect QTL of small effect (Odell et al., 2022). The advantages of DH populations are the capability of removing any residual heterozygosity to ensure genetically identical replicates and increasing selection response by stabilizing heritability of various traits during perse and test cross evaluation (Bordes et al., 2006; Gallais and Bordes, 2007; Mayor and Bernardo, 2009; Odell et al., 2022).

SNP markers are the most frequent variations in genomes and the application of SNP markers in plant breeding has guaranteed the precision of QTL mapping and genetic analysis (Bhattramakki et al., 2002; Mammadov et al., 2012; Flutre et al., 2022; Kaur et al., 2022). By conditioning linked markers in the test, the sensitivity of the test statistic to the position of individual QTLs is increased, and the precision of QTL mapping can be improved (Zeng, 1994). Subsequently, with the development of sequencing technology, an increasing number of molecular markers have been applied to QTL mapping, which greatly improves the accuracy of QTL mapping (Schnable et al., 2009; Chia et al., 2012; Bukowski et al., 2018; Fang et al., 2021). In this study, a total of 16 QTLs were found and distributed across all ten chromosomes. 13 QTLs spanned physical intervals of less than 50 Mb, and two span less than 10 Mb. Thus, the resolution in this study is considerably improved because of the large number of markers and the appropriate population type. The resolution is probably on the order of 2-3 cM, since pairs of markers any farther apart rarely have substantial levels of linkage disequilibrium (Laurie et al., 2004).

4.2 Genetic basis of TOC in our DH populations

Within the four DH populations, a broad range of phenotypic variation with normal distribution was observed for TOC with transgressive segregation, indicating quantitative genetic control (Figure 1). The identification of loci controlling oil-related traits should contribute to a better understanding of oil synthesis and storage in maize kernels. The genetic analysis indicated TOC is highly heritable and the heritability (67.00-86.60%) is fairly high in all populations, indicating of superior genetic effect on TOC in DH populations. The high heritability estimates are very favorable for detecting marker-trait associations (Laurie et al., 2004). Among the 16 detected QTLs controlling TOC, 11 QTLs were identified as the major QTLs with the explaining phenotypic variation larger than 10%. Especially *qOC-1-3* with the largest effect (30.84% of the phenotypic variance) and *qOC-2-3* with the second largest effect (21.74% of the phenotypic variance) were located on chromosome 9. These region have been chosen as our primary QTL for further study because of the higher contribution. The parent allele at this locus had an additive effect of 0.20-0.24% for increased TOC. An additional seven QTLs were identified on chromosomes 2, 4, 5, 6 and 8, explaining between 11.64 and 16.20% of the phenotypic variation. The other minor QTLs each explained 3.49-8.39% of the phenotypic variance with moderate additive effects on TOC. In addition, except for environment variation, none of QTLs were shared by all DH populations, reflecting the complexity of TOC regulation in diverse maize populations. These results indicated that oil content is controlled by a few large-effect QTLs, together with a large number of minor-effect QTLs (Dudley, 1977; Laurie et al., 2004).

Results of QTL detection derived from different studies may exhibit consistency to a certain degree across different germplasms or genetic backgrounds and environments. For instance, the largest and second effective QTL *qOC-1-3* and *qOC-2-3* was located in the QTL *m240* with a 29.17 Mb and 8.81 Mb overlap interval length, respectively, which was related to maize TOC in RIL population (Cook et al., 2012). *qOC-3-4* co-localized with *koc5b* associated to the kernel oil content in a $F_{2:3}$ tropical maize population (Mangolin et al., 2004). According to Li et al. (2013), the QTL *qOC-2-1*, *qOC-3-1*, *qOC-3-3*, *qOC-4-1* and *qOC-4-1* more or less co-localized with the QTLs controlling protein and TOC simultaneously and might affect protein and TOC in opposite directions (Li et al., 2013). These results suggested that increases in grain TOC might be associated with increases in grain protein content, both traits could be improved simultaneously. Congruence in QTLs detected in this study with previous reports indicates the robustness of our results. Moreover, these QTLs definitely worth conducting further research on this QTL *via* NILs, fine mapping, molecular marker-assisted selection (MAS) and ultimate cloning.

4.3 Importance of QTLs relevant to TOC in maize genetic and breeding

Oil in maize kernels mainly exists in the form of triacylglycerol (TAG), which composed of fatty acids and glycerol (Du et al., 2016; Zhang et al., 2019). Maize oil mainly accumulates in the embryo, and the fatty acids are typically comprised of approximately 11% palmitic acid (C16:0), 2% stearic acid (C18:0), 24% oleic acid (C18:1), 62% linoleic acid (C18:2), and 1% linolenic acid (C18:3) (Lambert, 2001). The quality and utilization of maize oil is determined by their fatty acid composition (Du et al., 2016). Saturated fatty acids, such as palmitic (C16:0) and stearic acids (C18:0), are stable and tolerant to heat and oxidation (Hu et al., 1997). Certain unsaturated fatty acids, such as oleic (C18:1), linoleic (C18:2), and linolenic (C18:3) acids, are beneficial to human health but susceptible to heat and oxidation (Hu et al., 1997). Biosynthesis of storage oil in plant seeds is complex and involved in multitudinous physiological and biochemical processes (Ohlrogge and Browse, 1995; Liu et al., 2008; Zhang et al., 2009; Guo et al., 2013; Dong et al., 2015; Glowinski and Flint-Garcia, 2018; Zhang et al., 2018). The co-location analysis of candidate genes underlying QTLs associated with related trait could provide information about functional relationships between gene expression and some QTLs of the complex biosynthesis pathway (Prioul et al., 1997; Thévenot et al., 2005). In our study, of 189 genes involved in the fatty acid biochemical processes, including 17 well-known genes encoding key enzymes in maize lipid synthesis and metabolism, were located within QTL intervals (Figure 4 and Table S2).

The genes related to the TAG synthesis pathway are key regulatory factors in the accumulation process of TOC in corn

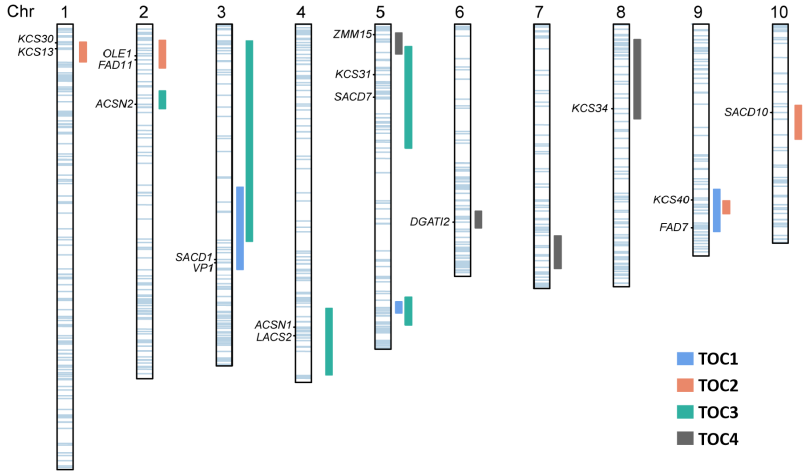


FIGURE 4
Association of candidate genes with kernel oil QTLs. The QTLs identified in four DH populations are represented as vertical rectangles of different colors next to each chromosome. The horizontal light blue bars on each chromosome show the positions of the 189 identified genes. The left labels denote known genes that co-localized with the QTLs.

(Zhang et al., 2019). Comparison of the positions of candidate genes and QTL was a suitable strategy to investigate the molecular basis of quantitative traits. Additionally, the positioned candidate genes can be used to develop functional markers for increasing selection efficiency by marker-assisted selection in plant breeding (Andersen and Lübberstedt, 2003). Five KCS genes encoding β-ketoacyl CoA synthase isozymes in *qOC-1-3*, *qOC-2-1*, *qOC-3-4* and *qOC-4-4* are mainly involved in the process of elongation of the C16:0- and C18:0-CoAs into very-long-chain fatty acids (VLCFAs) (Gonzales-Vigil et al., 2017). The maize isozymes reflected differences in the enzymatic capability to elongate fatty acids (Stenback et al., 2022). The *FAD* genes in *qOC-1-3* and *qOC-2-2* were identified as fatty acid desaturase-coding and are responsible for the production of trienoic fatty acids by unsaturation at the ω-3 position and the cDNAs corresponding to the loci have been isolated (Ohlrogge and Browse, 1995; Gao et al., 2015; Zhao et al., 2019). Stearoyl-acyl carrier protein desaturases (SACD) encoded by the genes in *qOC-1-1*, *qOC-2-4* and *qOC-3-4* are the key enzymes that converts stearic acid to oleic acid by introducing the first double bond into stearoyl-ACP between carbons 9 and 10 (Asamizu et al., 1998; Liu et al., 2009). These enzymes are significantly more abundant in expression in high-oil maize than in normal maize, not only at the mRNA and protein levels, but also at the product level (Liu et al., 2009). *LACS2* in *qOC-3-3* encoded the long-chain acyl-CoA synthetase (LACS), which plays key roles in activating fatty acids to fatty acyl-CoA thioesters and then further involved in lipid synthesis and fatty acid catabolism (Lü et al., 2009; Zhao et al., 2010; Jessen et al., 2011 and Jessen et al., 2015). TAG biosynthesis involves three consequential acylation steps of a glycerol backbone *via* the Kennedy pathway (Ohlrogge and Browse, 1995; Iskandarov et al., 2017; Müller and Ischebeck, 2018). The process starts with the acylation of glycerol-3-phosphate (G3P) by glycerol-3-phosphate acyltransferase (GPAT) and lysophosphatidic acid acyltransferase (LPAAT), and finalized by diacylglycerol acyltransferase (DGAT), which catalyzes the last acylation step of the pathway (Ohlrogge and Browse, 1995). The high-oil QTL (*qHO6*) affecting maize seed oil and oleic-acid contents encodes DGAT1-2 (Zheng et al., 2008; Yang et al., 2010; Hao et al., 2014). The gene *GPAT12* in our study was also detected on chromosome 6 and showed 96% identities with *DGAT1-2* (Zm00001d036982), which indicated that GPAT12 may be one of DGAT isozymes. The seed oils are packaged in spherical intracellular oil bodies, which have a TAG matrix surrounded by a layer of phospholipids embedded with unique and abundant proteins termed oleosins (Lee and Huang, 1994). Oleosins interact with the surface phospholipids and matrix triacylglycerols to form a stable amphipathic layer on the surface of the oil body and possibly act as recognition signals for the binding of lipase during germination (Lee and Huang, 1994; Lee et al., 1995; Ting et al., 1996). It suggested that *OLE1* in *qOC-2-2* was an important gene that would facilitate lipase action during germination. The above analysis suggested that the QTLs in this study were related to a series of genes encoding key enzymes relevant to oil content and lipid metabolism. Especially, *qOC-4-2* contained a DGAT1-2 homologous protein coding gene and had no common region with *qHO6* which was the major oil content QTL (Cook et al., 2012). Therefore, these QTLs will pave a path to explore molecular markers and offer prospective routes to improve maize oil content through molecular marker-assisted selection in maize breeding program.

5 Conclusion

In this study, four DH populations were constructed for genetic analysis of kernel TOC and the TOC exhibited continuously and approximately normal distribution in all populations. Six major and ten minor effect QTLs were identified based on the genetic linkage map with LOD threshold of 3.00 and accounted for 3.49-30.84% of oil variation. The result was consistent with Yang et al., 2010 that OC in maize kernel is a complex quantitative trait and controlled by

a few large-effect QTLs and numerous minor QTLs. Besides, 17 well-known genes involved in fatty acid synthesis and metabolic pathway were located within QTL intervals. This information provides insight that will help to further understanding of genetic variation in TOC in maize kernels and will thus enhance the feasibility of cloning QTL, lay the foundation to explore candidate genes associated with maize kernel TOC.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://doi.org/10.6084/m9.figshare.22152380.v1.

Author contributions

HL and CZ conceived and designed the experiments. XZ, HG, HW, CD, JW, BP, and JL performed the research. MW analysed the data. XZ and MW wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by National Natural Science Foundation of China, grant number 32201798 and Heilongjiang Scientific Research Business Expenses Project of China, grant number CZKYF2023-1-C001.

Acknowledgments

We thank all members of our laboratories for helpful discussion and assistance during this research.

Conflict of interest

Author CZ is employed by Maize Yufeng Biotechnology LLC Beijing, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1174985/full#supplementary-material

References

Alrefai, R., Berke, T. G., and Rocheford, T. R. (1995). Quantitative trait locus analysis of fatty acid concentrations in maize. *Genome.* 38, 894–901. doi: 10.1139/g95-118

Andersen, J. R., and Lübberstedt, T. (2003). Functional markers in plants. *Trends Plant Sci.* 8, 554–560. doi: 10.1016/j.tplants.09.010

Asamizu, E., Sato, S., Kaneko, T., Nakamura, Y., Kotani, H., Miyajima, N., et al. (1998). Structural analysis of *Arabidopsis thaliana* chromosome 5. VIII. Sequence features of the regions of 1,081,958 bp covered by seventeen physically assigned P1 and TAC clones. *DNA Res.* 31, 379–391. doi: 10.1093/dnares/5.6.379

Benitez, J. A., Gernat, A. G., Murillo, J. G., and Araba, M. (1999). The use of high oil corn in broiler diets. *Poult. Sci.* 78, 861–865. doi: 10.1093/ps/78.6.861

Berke, T. G., and Rocheford, T. R. (1995). Quantitative trait loci for flowering, plant and ear height, and kernel traits in maize. *Crop Sci.* 35, 1542–1549. doi: 10.2135/cropsci1995.0011183X003500060004x

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J. C., et al. (2002). Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Mol. Biol.* 48, 539–547. doi: 10.1023/a:1014841612043

Bordes, J., Charmet, G., de Vaulx, R. D., Pollacsek, M., Beckert, M., and Gallais, A. (2006). Doubled haploid versus S1 family recurrent selection for testcross performance in a maize population. *Theor. Appl. Genet.* 112, 1063–1072. doi: 10.1007/s00122-006-0208-3

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Bukowski, R., Guo, X. S., Lu, Y. L., Zou, C., He, B., Rong, Z. Q., et al. (2018). Construction of the third-generation *Zea mays* haplotype map. *Gigascience* 7, 1–12. doi: 10.1093/gigascience/gix134

Chaikam, V., Molenaar, W., Melchinger, A. E., and Boddupalli, P. M. (2019). Doubled haploid technology for line development in maize: technical advances and prospects. *Theor. Appl. Genet.* 132, 3227–3243. doi: 10.1007/s00122-019-03433-x

Chen, S. B., and Hu, Z. (2017). Determination of corn fat based on NIRS and QPSO-LSSVM model. *Chem. Engineer.* 31, 30–35. doi: 10.16247/j.cnki.23-1171/tq.20170830

Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., et al. (2012). Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* 44, 803–807. doi: 10.1038/ng.2313

Churchill, G. A., and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963–971. doi: 10.1093/genetics/138.3.963

Clark, D., Dudley, J. W., Rocheford, T. R., and LeDeaux, J. R. (2006). Genetic analysis of corn kernel chemical composition in the random mated 10 generation of the cross of generations 70 of IHO × ILO. *Crop Sci.* 3, 373–379. doi: 10.2135/cropsci2005.06-0153

Cook, J. P., McMullen, M. D., Holland, J. B., Tian, F., Bradbury, P., Ross-Ibarra, J., et al. (2012). Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.* 158, 824–834. doi: 10.1104/pp.111.185033

Dong, Y. B., Zhang, Z. W., Shi, Q. L., Wang, Q. L., Zhou, Q., and Li, Y. L. (2015). QTL identification and meta-analysis for kernel composition traits across three generations in popcorn. *Euphytica* 204, 649–660. doi: 10.1007/s10681-015-1360-0

Du, H. W., Huang, M., Hu, J. Y., and Li, J. S. (2016). Modification of the fatty acid composition in *Arabidopsis* and maize seeds using a stearoyl-acyl carrier protein desaturase-1 (*ZmSAD1*) gene. *BMC Plant Biol.* 16, 137. doi: 10.1186/s12870-016-0827-z

Du, L., Yu, F., Zhang, H., Wang, B., Ma, K. J., Yu, C. P., et al. (2020). Genetic mapping of quantitative trait loci and a major locus for resistance to grey leaf spot in maize. *Theor. Appl. Genet.* 133, 2521–2533. doi: 10.1007/s00122-020-03614-z

Dudley, J. W. (1977). "Seventy-six generation of selection for oil and protein percentage in maize," in *Proceedings of international conference on quantitative genetics*. Ed. E. Pollak (Iowa state, Ames: Iowa State University Press, Ames), 459–473.

Dudley, J. W. (2008). Epistatic interactions in crosses of Illinois high oil × Illinois low oil and of Illinois high protein × Illinois low protein corn strains. *Crop Sci.* 48, 59–68. doi: 10.2135/cropsci2007.04.0242

Dudley, J. W., and Lambert, R. J. (2004). 100 generations of selection for oil and protein in corn. *Plant Breed. Rev.* 24, 79–110.

Dupont, J., White, P. J., Carpenter, M. P., Schaefer, E. J., Meydani, S. N., Elson, C. E., et al. (1990). Food uses and health effects of corn oil. *J. Am. Coll. Nutr.* 9, 438–470. doi: 10.1080/07315724.1990.10720403

Fang, H., Fu, X. Y., Ge, H. Q., Zhang, A. X., Shan, T. Y., Wang, Y. D., et al. (2021). Genetic basis of maize kernel oil-related traits revealed by high-density SNP markers in a recombinant inbred line population. *BMC Plant Biol.* 21, 344. doi: 10.1186/s12870-021-03089-0

Fang, H., Fu, X. Y., Wang, Y. B., Xu, J., Feng, H. Y., Li, W. Y., et al. (2020). Genetic basis of kernel nutritional traits during maize domestication and improvement. *Plant J.* 101, 278–292. doi: 10.1111/tpj.14539

Flutre, T., Le Cunff, L., Fodor, A., Launay, A., Romieu, C., Berger, G., et al. (2022). A genome-wide association and prediction study in grapevine deciphers the genetic architecture of multiple traits and identifies genes under many new QTLs. *G3 (Bethesda)* 29, jkac103. doi: 10.1093/g3journal/jkac103

Foiada, F., Westermeier, P., Kessel, B., Ouzunova, M., Wimmer, V., Mayerhofer, W., et al. (2015). Improving resistance to the European corn borer: a comprehensive study in elite maize using QTL mapping and genome-wide prediction. *Theor. Appl. Genet.* 128, 875–891. doi: 10.1007/s00122-015-2477-1

Gallais, A., and Bordes, J. (2007). The use of doubled haploids in recurrent selection and hybrid development in maize. *Crop Sci.* 47, S190–S201. doi: 10.2135/cropsci2007.04.0019IPBS

Gao, J. P., Wallis, J. G., and Browse, J. (2015). Mutations in the prokaryotic pathway rescue the fatty acid biosynthesis1 mutant in the cold. *Plant Physiol.* 169, 442–452. doi: 10.1104/pp.15.00931

Glowinski, A., and Flint-Garcia, S. (2018). "Germplasm resources for mapping quantitative traits in maize," in *The maize genome. compendium of plant genomes*. Ed. R. Tuberosa (New York, NYC: Springer Press), 143–159.

Goldman, I. L., Rocheford, T. R., and Dudley, J. W. (1994). Molecular markers associated with maize kernel oil concentration in an Illinois high protein × Illinois low protein cross. *Crop Sci.* 34, 908–915. doi: 10.2135/cropsci1994.0011183X003400040013x

Gonzales-Vigil, E., Hefer, C. A., von Loessl, M. E., La Mantia, J., and Mansfield, S. D. (2017). Exploiting natural variation to uncover an alkene biosynthetic enzyme in poplar. *Plant Cell* 29, 2000–2015. doi: 10.1105/tpc.17.00338

Guo, Z. F., Wang, H. W., Tao, J. J., Ren, Y. H., Xu, C., Wu, K. S., et al. (2019). Development of multiple snp marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. *Mol. Breed.* 39, 37. doi: 10.1007/s11032-019-0940-4

Guo, Y. Q., Yang, X. H., Chander, S., Yan, J. B., Zhang, J., Song, T. M., et al. (2013). Identification of unconditional and conditional QTL for oil, protein and oil content in maize. *Crop J.* 1, 34–42. doi: 10.1016/j.cj.2013.07.010

Han, Y., Parsons, C. M., and Alexander, D. E. (1987). Nutritive value of high oil for poultry. *Poult. Sci.* 66, 103–111. doi: 10.3382/ps.0660103

Hao, X. M., Li, X. W., Yang, X. H., and Li, J. S. (2014). Transferring a major QTL for oil content using marker-assisted backcrossing into an elite hybrid to increase the oil content in maize. *Mol. Breed.* 34, 739–748. doi: 10.1007/s11032-014-0071-x

Hou, Q. C., Zhang, T. Y., Sun, K. T., Yan, T. W., Wang, L. L., Lu, L., et al. (2022). Mining of potential gene resources for breeding nutritionally improved maize. *Plants (Basel).* 11, 627. doi: 10.3390/plants11050627

Hu, F. B., Stampfer, M. J., Manson, J. E., Rimm, E., Colditz, G. A., Rosner, B. A., et al. (1997). Dietary fat intake and the risk of coronary heart disease in women. *N. Engl. J. Med.* 337, 1491–1499. doi: 10.1056/NEJM199711203372102

Iskandarov, U., Silva, J. E., Kim, H. J., Andersson, M., Cahoon, R. E., Mockaitis, K., et al. (2017). A specialized diacylglycerol acyltransferase contributes to the extreme medium-chain fatty acid content of cuphea seed oil. *Plant Physiol.* 174, 97–109. doi: 10.1104/pp.16.01894

Jessen, D., Olbrich, A., Knüfer, J., Krüger, A., Hoppert, M., Polle, A., et al. (2011). Combined activity of LACS1 and LACS4 is required for proper pollen coat formation in *Arabidopsis*. *Plant J.* 68, 715–726. doi: 10.1111/j.1365-313X.2011.04722.x

Jessen, D., Roth, C., Wiermer, M., and Fulda, M. (2015). Two activities of long-chain acyl-coenzyme a synthetase are involved in lipid trafficking between the endoplasmic reticulum and the plastid in *Arabidopsis*. *Plant Physiol.* 167, 351–366. doi: 10.1104/pp.114.250365

Karn, A., Gillman, J. D., and Flint-Garcia, S. A. (2017). Genetic analysis of teosinte alleles for kernel composition traits in maize. *G3 (Bethesda).* 7, 1157–1164. doi: 10.1534/g3.117.039529

Kaur, G., Pathak, M., Singla, D., Chhabra, G., Chhuneja, P., and Kaur Sarao, N. (2022). Quantitative trait loci mapping for earliness, fruit, and seed related traits using high density genotyping-by-sequencing-based genetic map in bitter gourd (*Momordica charantia* l.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.799932

Knapp, S. J., Stroup, W. W., and Ross, W. M. (1985). Exact confidence-intervals for heritability on a progeny mean basis. *Crop Sci.* 25, 192–194. doi: 10.2135/cropsci1985.0011183X002500010046x

Lambert, R. J. (2001). "High-oil corn hybrids," in *Specialcorn*. Ed. A. R. Hallau (Boca Raton: CRC Press), 131–153.

Lambert, R. J., Alexander, D. E., and Mejaya, I. J. (2004). Single kernel selection for increased grain oil in maize synthetics and high-oil hybrid development. *Plant Breed. Rev.* 24, 153–175. doi: 10.1002/9780470650240.ch8

Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121, 185–199. doi: 10.1093/genetics/121.1.185

Laurie, C. C., Chasalow, S. D., LeDeaux, J. R., McCarroll, R., Bush, D., Hauge, B., et al. (2004). The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics* 168, 2141–2155. doi: 10.1534/genetics.104.029686

Lee, K. Y., and Huang, A. H. (1994). Genes encoding oleosins in maize kernel of inbreds Mo17 and B73. *Plant Mol. Biol.* 26, 1981–1987. doi: 10.1007/BF00019508

Lee, K. Y., Ratnayake, C., and Huang, A. H. (1995). Genetic dissection of the co-expression of genes encoding the two isoforms of oleosins in the oil bodies of maize kernel. *Plant J.* 7, 603–611. doi: 10.1046/j.1365-313x.1995.7040603.x

Li, L., Li, H., Li, Q., Yang, X. H., Zheng, D. B., Warburton, M., et al. (2011). An 11-bp insertion in *Zea mays fatb* reduces the palmitic acid content of fatty acids in maize grain. *PloS One* 6, e24699. doi: 10.1371/journal.pone.0024699

Li, H., Peng, Z. Y., Yang, X. H., Wang, W. D., Fu, J. J., Wang, J. H., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, H., Wang, M., Li, W. J., He, L. L., Zhou, Y. Y., Zhu, J. T., et al. (2020). Genetic variants and underlying mechanisms influencing variance heterogeneity in maize. *Plant J.* 103, 1089–1102. doi: 10.1111/tpj.14786

Lima, M., d., A., de Souza, C. L., Bento, D. A. V., Souza, A., and Carlini-Garcia, L. A. (2006). Mapping QTL for grain yield and plant traits in a tropical maize population. *Mol. Breed.* 17, 227–239. doi: 10.1007/s11032-005-5679-4

Liu, Y. Y., Dong, Y. B., Niu, S. Z., Cui, D. Q., Wang, Y. Z., Wei, M. G., et al. (2008). QTL identification of kernel composition traits with popcorn using both $F_{2:3}$ and BC_2F_2 populations developed from the same cross. *J. Cereal Sci.* 48, 625–631. doi: 10.1016/j.jcs.2008.02.003

Liu, Z. J., Yang, X. H., Fu, Y., Zhang, Y. R., Yan, J. B., Song, T. M., et al. (2009). Proteomic analysis of early germs with high-oil and normal inbred lines in maize. *Mol. Biol. Rep.* 6, 813–821. doi: 10.1007/s11033-008-9250-3

Lü, S. Y., Song, T., Kosma, D. K., Parsons, E. P., Rowland, O., and Jenks, M. A. (2009). *Arabidopsis CER8* encodes LONG-CHAIN ACYL-COA SYNTHETASE 1 (LACS1) that has overlapping functions with LACS2 in plant wax and cutin synthesis. *Plant J.* 59, 553–564. doi: 10.1111/j.1365-313X.2009.03892.x

Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *Int. J. Plant Genomics* 2012, 728398. doi: 10.1155/2012/728398

Mangolin, C. A., de Souza, C. L., Garcia, A. A. F., Garcia, A. F., Sibov, S. T., and de Souza, A. P. (2004). Mapping QTLs for kernel oil content in a tropical maize population. *Euphytica* 137, 251–259. doi: 10.1023/B:EUPH.0000041588.95689.47

Mayor, P. J., and Bernardo, R. (2009). Genomewide selection and marker-assisted recurrent selection in doubled haploid versus f populations. *Crop Sci.* 49, 1719–1725. doi: 10.2135/cropsci2008.10.0587

Messmer, R., Fracheboud, Y., Bänziger, M., Vargas, M., Stamp, P., and Ribaut, J. M. (2009). Drought stress and tropical maize: QTL-by-environment interactions and stability of QTLs across environments for yield components and secondary traits. *Theor. Appl. Genet.* 119, 913–930. doi: 10.1007/s00122-009-1099-x

Müller, A. O., and Ischebeck, T. (2018). Characterization of the enzymatic activity and physiological function of the lipid droplet-associated triacylglycerol lipase AtOBL1. *New Phytol.* 217, 1062–1076. doi: 10.1111/nph.14902

Odell, S. G., Hudson, A. I., Praud, S., Dubreuil, P., Tixier, M. H., Ross-Ibarra, J., et al. (2022). Modeling allelic diversity of multiparent mapping populations affects detection of quantitative trait loci. *G3 (Bethesda)* 12, jkac011. doi: 10.1093/g3journal/jkac011

Ohlrogge, J., and Browse, J. (1995). Lipid biosynthesis. *Plant Cell* 7, 957–970. doi: 10.1105/tpc.7.7.957

Pan, Q. C., Ali, F., Yang, X. H., Li, J. S., and Yan, J. B. (2012). Exploring the genetic characteristics of two recombinant inbred line populations *via* high-density SNP markers in maize. *PloS One* 7, e52777. doi: 10.1371/journal.pone.0052777

Prioul, J. L., Quarrie, S., Causse, M., and de Vienne, D. (1997). Dissecting complex physiological functions through the use of molecular quantitative genetics. *J. Exp. Bot.* 48, 1151–1163. doi: 10.1093/jxb/48.6.1151

Ranum, P., Peña-Rosas, J. P., and Garcia-Casal, M. N. (2014). Global maize production, utilization, and consumption. *Ann. N. Y. Acad. Sci.* 1312, 105–112. doi: 10.1111/nyas.12396

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F. S., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Song, T. M., and Chen, S. J. (2004). Long term selection for oil concentration in five maize populations. *Maydica* 49, 9–14. Available at: https://www.researchgate.net/publication/289088076_Long_term_selection_for_oil_concentration_in_five_maize_populations.

Stenback, K. E., Flyckt, K. S., Hoang, T., Campbell, A. A., and Nikolau, B. J. (2022). Modifying the yeast very long chain fatty acid biosynthetic machinery by the expression of plant 3-ketoacyl CoA synthase isozymes. *Sci. Rep.* 12, 13235. doi: 10.1038/s41598-022-17080-8

Thévenot, C., Simond-Côte, E., Reyss, A., Manicacci, D., Trouverie, J., Le Guilloux, M., et al. (2005). QTLs for enzyme activities and soluble carbohydrates involved in oil

accumulation during grain filling in maize. *J. Exp. Bot.* 56, 945–958. doi: 10.1093/jxb/eri087

Ting, J. T., Lee, K., Ratnayake, C., Platt, K. A., Balsamo, R. A., and Huang, A. H. (1996). Oleosin genes in maize kernels having diverse oil contents are constitutively expressed independent of oil contents. size and shape of intracellular oil bodies are determined by the oleosins/oils ratio. *Planta* 199, 158–165. doi: 10.1007/BF00196892

Wang, S., Basten, C. J., and Zeng, Z. B. (2010a). *Windows QTL cartographer V2.5_011* (North Carolina state, Raleigh: Dep. Stat. North Carolina State University).

Wang, Y. Z., Li, J. Z., Li, Y. L., Wei, M. G., and Fu, J. F. (2010b). QTL detection for grain oil and oil content and their associations in two connected $F_{2:3}$ populations in high-oil maize. *Euphytica* 174, 239–252. doi: 10.1007/s10681-010-0123-1

Wang, T. T., Wang, M., Hu, S. T., Xiao, Y. N., Tong, H., Pan, Q. C., et al. (2015). Genetic basis of maize kernel starch content revealed by high-density single nucleotide polymorphism markers in a recombinant inbred line population. *BMC Plant Biol.* 15, 288. doi: 10.1186/s12870-015-0675-2

Wassom, J. J., Mikkelineni, V., Bohn, M. O., and Rocheford, T. R. (2008a). QTL for fatty acid composition of maize kernel oil in Illinois high oil x B73 backcross-derived lines. *Crop Sci.* 48, 69–78. doi: 10.2135/cropsci2007.04.0208

Wassom, J. J., Wong, J. C., Martinez, E., King, J. J., Debaene, J., Hotchkiss, J. R., et al. (2008b). QTL associated with maize kernel oil, protein, and oil concentrations; kernel mass; and grain yield in illinois high oil × B73 backcross-derived lines. *Crop Sci.* 48, 243–252. doi: 10.2135/cropsci2007.04.0205

Wei, M., Fu, J., Li, X., Wang, Y., and Li, Y. (2009). Influence of dent corn genetic backgrounds on QTL detection for plant-height traits and their relationships in high-oil maize. *J. Appl. Genet.* 50, 225–234. doi: 10.1007/BF03195676

Yan, G. J., Liu, H., Wang, H. B., Lu, Z. Y., Wang, Y. X., Mullan, D., et al. (2017). Accelerated generation of selfed pure line plants for gene identification and crop breeding. *Front. Plant Sci.* 8, 1786. doi: 10.3389/fpls.2017.01786

Yang, X. H., Guo, Y. Q., Yan, J. B., Zhang, J., Song, T. M., Rocheford, T., et al. (2010). Major and minor QTL and epistasis contribute to fatty acid compositions and oil concentration in high-oil maize. *Theor. Appl. Genet.* 120, 665–678. doi: 10.1007/s00122-009-1184-1

Yang, Z., Li, X., Zhang, N., Zhang, Y. N., Jiang, H. W., Gao, J., et al. (2016). Detection of quantitative trait loci for kernel oil and protein concentration in a B73 and Zheng58 maize cross. *Genet. Mol. Res.* 15, 1–10. doi: 10.4238/gmr.15038951

Yang, X. H., Ma, H. L., Zhang, P., Yan, J. B., Guo, Y. Q., Song, T. M., et al. (2012). Characterization of QTL for oil content in maize kernel. *Theor. Appl. Genet.* 125, 1169–1179. doi: 10.1007/s00122-012-1903-x

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136, 1457–1468. doi: 10.1093/genetics/136.4.1457

Zhang, M., Fan, J. L., Taylor, D. C., and Ohlrogge, J. B. (2009). *DGAT1* and *PDAT1* acyltransferases have overlapping functions in *Arabidopsis* triacylglycerol biosynthesis and are essential for normal pollen and seed development. *Plant Cell.* 21, 3885–3901. doi: 10.1105/tpc.109.071795

Zhang, Y. H., He, J. B., Wang, H. W., Meng, S., Xing, G. N., Li, Y., et al. (2018). Detecting the QTL-allele system of seed oil traits using multi-locus genome-wide association analysis for population characterization and optimal cross prediction in soybean. *Front. Plant Sci.* 9, 1793. doi: 10.3389/fpls.2018.01793

Zhang, X. X., Hong, M. Y., Wan, H. P., Luo, L. X., Yu, Z., and Guo, R. X. (2019). Identification of key genes involved in embryo development and differential oil accumulation in two contrasting maize genotypes. *Genes (Basel)* 10, 993. doi: 10.3390/genes10120993

Zhang, J., Lu, X. Q., Song, X. F., Yan, J. B., Song, T. M., Dai, J. R., et al. (2008). Mapping quantitative trait loci for oil, oil, and protein concentrations in grain with high-oil maize by SSR markers. *Euphytica* 162, 335–344. doi: 10.1007/s10681-007-9500-9

Zhang, X. L., Lu, M., Xia, A. A., Xu, T., Cui, Z. H., Zhang, R. Y., et al. (2021). Genetic analysis of three maize husk traits by QTL mapping in a maize-teosinte population. *BMC Genomics* 22, 386. doi: 10.1186/s12864-021-07723-x

Zhang, X. L., Wang, M., Zhang, C. Z., Dai, C. J., Guan, H. T., and Zhang, R. Y. (2022). Genetic dissection of QTLs for starch content in four maize DH populations. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.950664

Zhao, L. F., Katavic, V., Li, F. L., Haughn, G. W., and Kunst, L. (2010). Insertional mutant analysis reveals that long-chain acyl-CoA synthetase 1 (LACS1), but not LACS8, functionally overlaps with LACS9 in *Arabidopsis* seed oil biosynthesis. *Plant J.* 64, 1048–1058. doi: 10.1111/j.1365-313X.2010.04396.x

Zhao, X. C., Wei, J. P., He, L., Zhang, Y. F., Zhao, Y., Xu, X. X., et al. (2019). Identification of fatty acid desaturases in maize and their differential responses to low and high temperature. *Genes (Basel)* 10, 445. doi: 10.3390/genes10060445

Zheng, P. Z., Allen, W. B., Roesler, K., Williams, M. E., Zhang, S. R., Li, J. M., et al. (2008). A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat. Genet.* 40, 367–372. doi: 10.1038/ng.85

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 May 2023
DOI 10.3389/fpls.2023.1189456

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Xingwang Yu,
North Carolina State University, United States
Jianfang Li,
Guangzhou Laboratory, China

*CORRESPONDENCE
Jingjing Qian
qianjingjing19@126.com
Yuchen Yang
yangych68@mail.sysu.edu.cn

RECEIVED 19 March 2023
ACCEPTED 17 April 2023
PUBLISHED 08 May 2023

CITATION
Xia X, Fan M, Liu Y, Chang X, Wang J, Qian J and Yang Y (2023) Genome-wide alternative polyadenylation dynamics underlying plant growth retardant-induced dwarfing of pomegranate. *Front. Plant Sci.* 14:1189456. doi: 10.3389/fpls.2023.1189456

Genome-wide alternative polyadenylation dynamics underlying plant growth retardant-induced dwarfing of pomegranate

Xinhui Xia[1], Minhong Fan[1], Yuqi Liu[1], Xinyue Chang[1], Jingting Wang[2], Jingjing Qian[2]* and Yuchen Yang[1]*

[1]State Key Laboratory of Biocontrol, School of Ecology, Sun Yat-sen University, Shenzhen, China, [2]College of Agriculture, Anhui Science and Technology University, Fengyang, China

Dwarfed stature is a desired agronomic trait for pomegranate (*Punica granatum* L.), with its advantages such as lower cost and increased yield. A comprehensive understanding of regulatory mechanisms underlying the growth repression would provide a genetic foundation to molecular-assisted dwarfing cultivation of pomegranate. Our previous study induced dwarfed pomegranate seedlings *via* exogenous application of plant growth retardants (PGRs) and highlighted the important roles of differential expression of plant growth-related genes in eliciting the dwarfed phenotype of pomegranate. Alternative polyadenylation (APA) is an important post-transcriptional mechanism and has been demonstrated to act as a key regulator in plant growth and development. However, no attention has been paid to the role of APA in PGR-induced dwarfing in pomegranate. In this study, we characterized and compared APA-mediated regulation events underlying PGR-induced treatments and normal growth condition. Genome-wide alterations in the usage of poly(A) sites were elicited by PGR treatments, and these changes were involved in modulating the growth and development of pomegranate seedlings. Importantly, ample specificities were observed in APA dynamics among the different PGR treatments, which mirrors their distinct nature. Despite the asynchrony between APA events and differential gene expression, APA was found to regulate transcriptome *via* influencing microRNA (miRNA)-mediated mRNA cleavage or translation inhibition. A global preference for lengthening of 3′ untranslated regions (3′ UTRs) was observed under PGR treatments, which was likely to host more miRNA target sites in 3′ UTRs and thus suppress the expression of the corresponding genes, especially those associated with developmental growth, lateral root branching, and maintenance of shoot apical meristem. Together, these results highlighted the key role of APA-mediated regulations in fine-tuning the PGR-induced dwarfed stature of pomegranate, which provides new insights into the genetic basis underlying the growth and development of pomegranate.

KEYWORDS

alternative polyadenylation, plant growth retardant, pomegranate, post-transcriptional regulation, dwarfing

Introduction

Pomegranate (*Punica granatum* L.) is one type of the economic fruit trees that are widely cultivated across the globe. Because it is rich in vitamins and has antioxidant and anti-inflammatory properties in fruits, the health benefits of pomegranate are highly regarded, such as preventing or alleviating diseases and lowering high blood pressure or high cholesterol levels (National Center for Complementary and Integrative Health, NCCIH; Bourekoua et al., 2018; Shahamirian et al., 2019; Asrey et al., 2020; Turrini et al., 2020). With the fast-rising demand for pomegranate products, more and more attention has been paid to screen and breed pomegranate cultivars with the desired high fruit yield and quality. Dwarfing cultivation is one of the major focuses because of its advantages in plant photosynthetic efficiency, fruit production, and disease resistance compared to normal growing mode (Seleznyova et al., 2008; Foster et al., 2017; Wang et al., 2018; Zhou and Underhill, 2021). Qian et al. (2022) demonstrated that exogenous application of plant growth retardants (PGRs) can successfully elicit dwarfed pomegranate seedlings. Comparative transcriptome analysis further unraveled that PGR-mediated downregulation of plant growth hormone synthesis played a central role in inducing the dwarfed stature of pomegranate, providing new clues for molecular breeding of favorable dwarfed pomegranate varieties. Besides gene transcription, plant transcriptome is also under the regulation of post-transcriptional mechanisms, which have been demonstrated as a key contributor to the phenotypic plasticity of plants (Ye et al., 2019; Zhou et al., 2019; Singh and Roychoudhury, 2021). However, our current knowledge on the functional importance of post-transcriptional processes in pomegranate is still limited.

Polyadenylation [poly(A)] is an important post-transcriptional mechanism in eukaryotes that modulates mRNA maturation from the precursor mRNA (pre-mRNA). It includes two coupled steps: endonucleolytic cleavage at the 3' end of pre-mRNA and the addition of a poly(A) tail at the cleavage sites (Colgan and Manley, 1997; Tian and Manley, 2017). More importantly, for many genes, the cleavage and poly(A) signal recognition occur at multiple positions, that is, giving rise to multiple mRNA isoforms with different lengths, which is referred to as alternative polyadenylation (APA). APA events have been demonstrated to be widespread across genomes; for example, over 70% of the *Arabidopsis* genes were found to possess more than one poly(A) site (Wu et al., 2011; Elkon et al., 2013). These APA events may alter the stability and translation of mRNA or the length of the resulting protein products; thus, APA serves as a key contributor to the complexity of eukaryotic transcriptome (Shen et al., 2008; Di Giammartino et al., 2011; Sun et al., 2012; Tian and Manley, 2017). Recent studies have highlighted the biological importance of APA in regulating plant growth, development, and resistance to environmental stresses (de Lorenzo et al., 2017; Zhou et al., 2019; Yu et al., 2022; Wang et al., 2023). For instance, Yu et al. (2022) performed a genome-wide investigation to APA dynamics underlying *Arabidopsis* leaf ontogeny and showed that the largest changes in poly(A) site usage occurred at the early stage of true leaf development, while the APA levels experienced a reduction along the developmental process. Furthermore, it was shown that these APA genes participated in modulating the biological processes associated with leaf development, for example, response to phytohormone. These findings highlighted the essential roles of APA-mediated post-transcriptional regulations in plant growth and development. However, the APA mechanisms underlying PGR-induced dwarfing have not been investigated in pomegranate.

In this study, we reanalyzed the published RNA-seq datasets (Qian et al., 2020) and characterized the genome-wide APA dynamics in the pomegranate seedlings treated with three kinds of PGRs, paclobutrazol, B9, and mannitol, to decipher the biological significance of APA-mediated mechanisms underlying PGR-induced dwarfing in pomegranate. Furthermore, we also compared the APA regulation to the gene expression changes, with the aim of dissecting the different contributions of transcriptional and post-transcriptional mechanisms to growth repression in pomegranate. Our findings will broaden our understanding of the genetic basis behind the PGR-elicited dwarfed stature of pomegranate and provide a foundation for future molecular-assisted dwarfing cultivation of pomegranate.

Materials and methods

Plant materials and data preprocessing

In our previous study, gene expression was characterized for the seedlings untreated (control group, CK) and treated with each of the three PGRs at different concentrations (paclobutrazol: 6 and 8 mg/L; B9: 6 and 8 mg/L; mannitol: 2.5 and 15 g/L) (Qian et al., 2022). Here, we reanalyzed the 14 RNA-seq datasets (two biological replicates for each scenario), which were deposited in the Gene Expression Omnibus (GEO) database of the National Center for Biotechnology Information (NCBI) under the accession number GSE195722, to investigate genome-wide poly(A) usage dynamics under the PGR treatments over CK. Data preprocessing was performed following the pipelines described in Qian et al. (2022). Briefly, for each dataset, low-quality bases, whose quality score < 20, and adapter contamination were first trimmed from the end of reads using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Simultaneously, the reads with either error rate > 0.1 or ambiguous/N bases > 15 were discarded from the dataset. Finally, the sequences with length after trimming < 50 bp were also excluded from the downstream analysis. The clean reads were mapped to the pomegranate reference genome (the soft-seeded pomegranate cultivar "Tunisia") *via* HISAT2 (Kim et al., 2019; Luo et al., 2020). The reads uniquely aligned to the genome were extracted and converted into bedgraph format using the sub-command *genomecov* of the BEDTools suite for downstream analysis (Quinlan and Hall, 2010).

Differentially expressed alternative polyadenylation analysis

The gene model file for the reference genome in 12-column bed format (bed12) was converted from the GTF-format genome

annotations file using the UCSC tools, *gtfToGenePred* and *genePredToBed* (https://genome.ucsc.edu/). The alignment result in bedgraph format and the gene model file were used as the inputs for the APA dynamics analysis using the APAtrap toolkit (Ye et al., 2018). Specifically, the annotated 3' untranslated regions (3' UTRs) were first refined and novel 3' UTRs or 3' UTR extensions were detected based on the mapping results of all the samples by the *identifyDistal3UTR* program. All the putative APA sites, as well as the usage level of APA sites, were predicted using *predictAPA* with default parameter settings. Differential usage analysis was performed for APA sites between CK and each of the treatment scenarios using the R package *deAPA*. The genes with an adjusted *p*-value < 0.05 and percentage difference (PD) ≥ 0.1 were considered to be significantly different in APA site usage between two groups, which were denoted as "differentially expressed APA genes (DAGs)". The functional importance of the DAGs was assessed by Gene Ontology (GO) enrichment analysis using Fisher's exact test, where the GO terms with *p*-value < 0.05 were considered to be significantly overrepresented compared to the genome background.

Prediction of putative microRNA target sites

The majority of APA events occur in 3' UTRs, that is, producing mRNA isoforms with 3' UTRs of different lengths. Changes in the length of 3' UTRs may cause the presence or loss of cis-regulatory elements, and thus pose influences on the stability, nuclear export, and translation efficiency of mRNA (Shen et al., 2008; Di Giammartino et al., 2011; Sun et al., 2012; Tian and Manley, 2017). Here, for each comparison, the DAGs were first grouped into two categories based on the Pearson product moment correlation coefficient *r*: (1) DAGs with $r < 0$ were supposed to contain more abundant proximal poly(A) site/shortened 3' UTR under the treatment than CK, while (2) DAGs with $r > 0$ were indicated to use more distal poly(A) site/lengthened 3' UTR in the treatment scenario. For each DAG of each category, the DNA sequence of each APA isoform was extracted from the reference genome using the sub-command *fastaFromBed* of BEDTools, and the putative microRNA (miRNA) target sites were identified in the 3' UTR by screening against the collected miRNA sequences in miRBase (Release 21) using the psRNATarget web server (http://plantgrn.noble.org/psRNATarget/). The maximum cutoff of complementary matching score was set to 4.0. The isoforms undergoing 3' UTR lengthening were supposed to be under the extra regulation of the miRNAs whose target sites were located in the lengthened 3' end, compared to those with shorter 3' UTRs.

Comparison between differentially expressed alternative polyadenylation and differentially expressed genes

To further investigate the different regulatory roles of gene transcription and APA in PGR-induced dwarfing, we compared the DAGs to the differentially expressed genes (DEGs) detected in our previous study (Qian et al., 2022). The overlapping between DAGs and DEGs was visualized by a Venn diagram using the *draw.pairwise.venn* function of the R package VennDiagram. GO enrichment analysis was implemented with a cutoff *p*-value of 0.05 for the genes from each of the three categories: (1) the genes under the regulation of both differential expression and APA; (2) the genes specifically regulated by differentially expressed APA (DA-specific genes); and (3) the genes specifically regulated by differential expression (DE-specific genes).

Results

PGR-induced alternative polyadenylation changes play a substantial role in regulating pomegranate growth

Compared to CK, exogenous applications of PGRs elicited 289–2,553 DAGs with significant differentiations in APA usage (Figure 1A). Functional enrichment analysis showed that these PGR-responsive APA events were associated with the biological processes of plant growth and development (Figures 1B–D). For instance, the DAGs induced by 8 mg/L B9 were enriched in auxin transport, root development, and maintenance of shoot apical meristem identity (Figure 1B), and mannitol-responsive DAGs were predominantly involved in the GO terms of leaf development and senescence, photomorphogenesis, stomatal movement, and cellular response to osmotic stress (Figure 1C). The application of 6 mg/L paclobutrazol was found to affect growth regulation and cell wall biosynthesis, and DAGs under the 8 mg/L treatment was overrepresented in leaf development (Figure 1D).

For all the PGRs, the treatment at high concentration could provoke more alterations in APA profiles than that at low level (Figure 1A), which was consistent with their larger effects on suppressing the growth of pomegranate seedlings (Qian et al., 2022). With regard to paclobutrazol, 341 DAGs were shared between the treatments at the two concentrations, which were overrepresented in the regulations of growth rate and leaf senescence (Figure 2A). Comparatively, 241 and 2,212 genes displayed 3' UTR alterations specifically under 6 and 8 mg/L treatment, respectively (Figure 2A). In particular, the DAGs specifically induced by 6 mg/L paclobutrazol were enriched in leaf development and thylakoid, whereas those responsive to 8 mg/L treatment were involved in cell tip growth and stomatal movement regulation (Figure 2A). Similar concentration-level specificities were also observed in the treatments of B9 and mannitol (Supplementary Figure 1). For instance, the 6 mg/L B9 treatment altered the poly(A) site usage of the genes related to seedling development, shoot apical meristem development, and cell wall thickening, and the DAGs identified under the 8 mg/L treatment were overrepresented in auxin transport, developmental process, and maintenance of shoot apical meristem identity (Supplementary Figure 1B, left panel). Together, these functional specificities of APA events unraveled the dose–response relationships of PGR treatments, which may assist to determine the optimal concentration for PGR application.

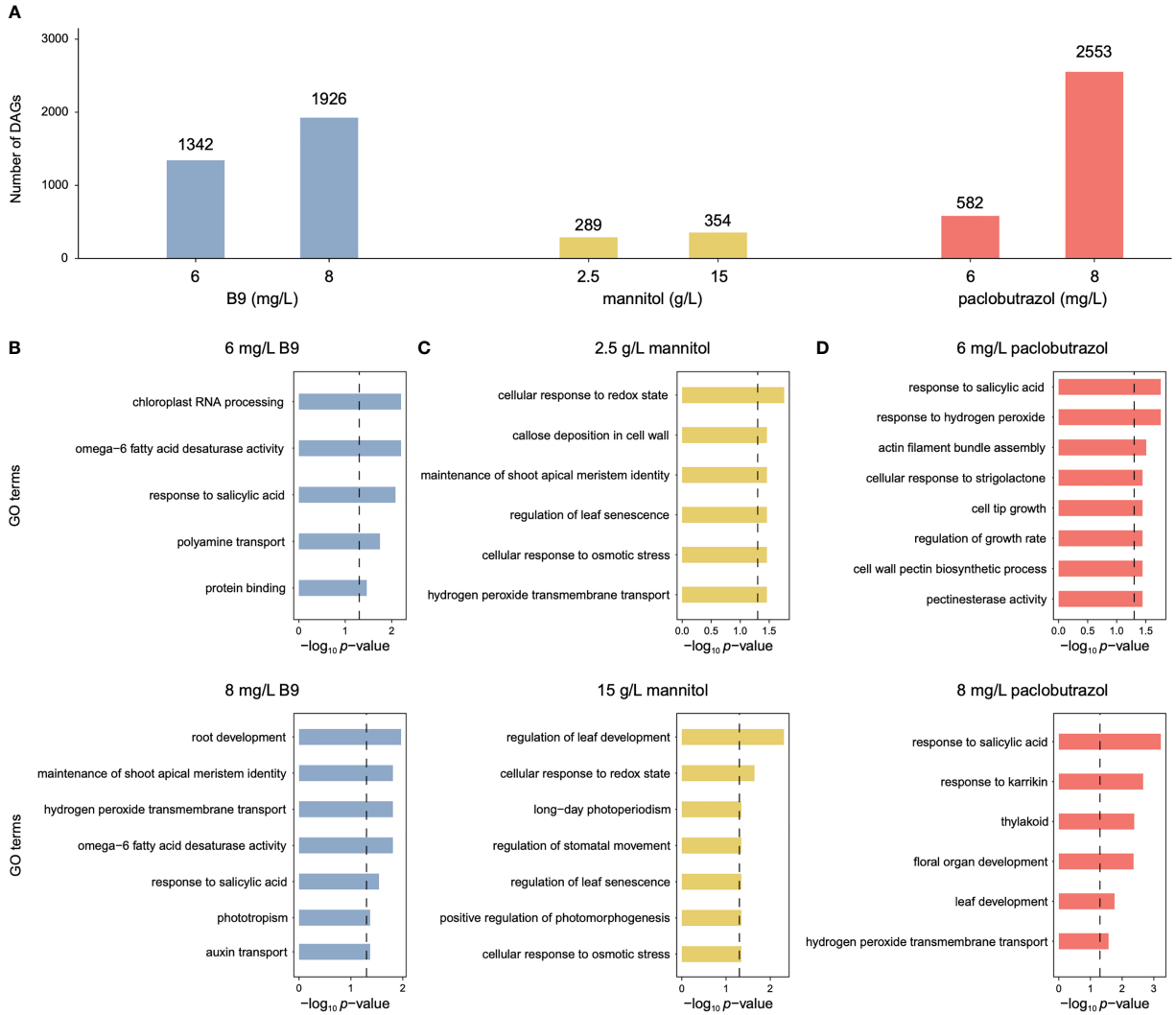


FIGURE 1
Numbers of DAGs induced by three PGRs and the corresponding enriched GO terms. **(A)** Numbers of DAGS identified under the treatment of B9 (blue), mannitol (yellow), and paclobutrazol (red). **(B–D)** Featur GO terms significantly enriched for the DAGs responsive to B9 **(B)**, mannitol **(C)**, and paclobutrazol **(D)**.

Different alternative polyadenylation regulations were elicited by different PGRs

The APA changes also showed substantial specificities among different PGR treatments (Figure 2B). In total, 1,308, 693, and 109 APA events occurred exclusively when exogenously applied with 8 mg/L paclobutrazol, 8 mg/L B9, and 15 g/L mannitol, respectively, while only 147 events were observed in all these three treatments. Functional enrichment analysis showed that, the commonly changed events were supposed to mainly affect leaf development and senescence, stomatal closure, hydrogen peroxide transmembrane transport, and cellular response to redox state (Figure 2B). Comparatively, the DAGs specifically elicited by 8 mg/L paclobutrazol were enriched in the GO terms of mitochondrial respiratory chain complex I, thylakoid, cellulose biosynthesis, and karrikin response, while those exclusively occurring under the 8 mg/L B9 treatment were overrepresented in auxin transport, xanthophyll biosynthesis, and phototropism (Figure 2B). The biological processes involved in leaf development, tricarboxylic acid transmembrane transport, signal transduction, and hydrogen peroxide biosynthesis were enriched for the specially induced DAGs by 15 g/L mannitol (Figure 2B).

Variations in the expression level of core polyadenylation factors, including polyadenylation machinery components, RNA-binding proteins, and transcription-related process, have been found to regulate APA (Hunt et al., 2012; Tian and Manley, 2017). Here, we first identified the genes encoding the subunits of four types of plant polyadenylation factors, cleavage stimulatory factor (CstF), cleavage and polyadenylation specificity factor (CPSF), poly(A) binding proteins (PABPs), and factor interacting with poly(A) polymerase (FIP1), in the pomegranate genome and compared their expression profiles under each treatment scenario to CK. The results showed that there were several polyadenylation

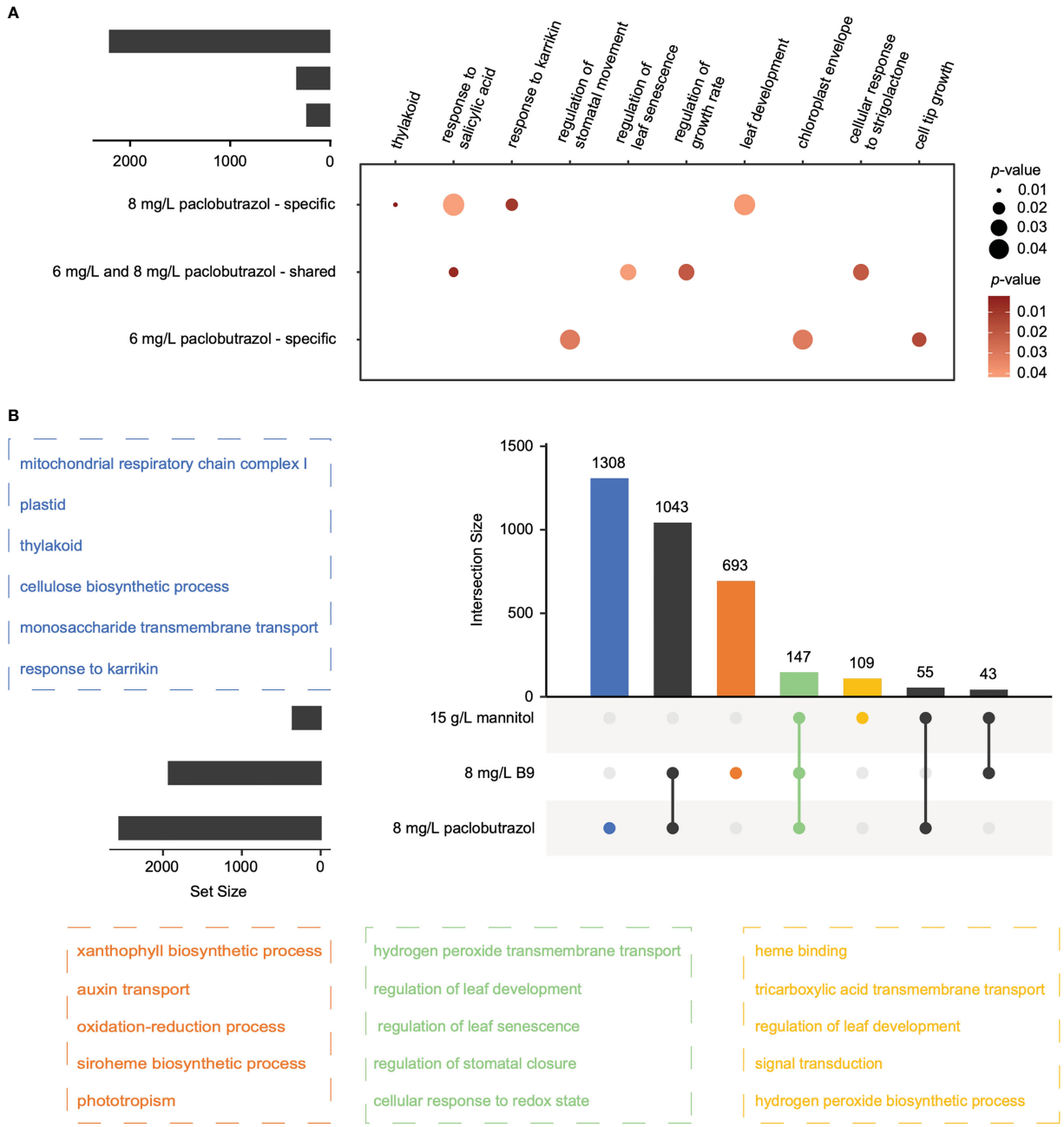


FIGURE 2
Overlap of the DAGs across treatments. **(A)** DAGs overlapped between 6 and 8 mg/L paclobutrazol treatments, and the representative GO terms enriched for each category. Circle size and color represent the significance level (*p*-value) of enrichment. **(B)** DAGs overlapped among the treatments of 8 mg/L paclobutrazol, 8 mg/L B9, and 15 g/L mannitol, and the representative GO terms enriched for the DAGs of different categories (highlighted by different colors).

factors significantly differentially expressed in response to the application of PGRs, which may play an important role in APA regulations (Figure 3). It is noteworthy that the expression profiles of polyadenylation apparatuses also displayed ample specificities among different treatments. Only one factor, PABP 7B, was commonly differentially expressed across all the three PGRs. In contrast, the expression of PABP 5 and 6 was specifically altered under the treatment of B9, and the gene that encodes PABP 9 and CPSF subunit 6 had an exclusively differential expression when treated with 8 mg/L paclobutrazol (Figure 3).

Alternative polyadenylation and gene transcription variations play a relatively independent role in growth regulation

To explore the different roles of gene transcription and APA dynamics in the regulation of PGR-induced dwarfing in pomegranate, we compared DAGs identified in each treatment to DEGs of the corresponding scenario. When treated with PGRs, most genes were specifically under the regulation of either gene expression or APA (Figure 4 and Supplementary Figures 2, 3). For instance, 24 and

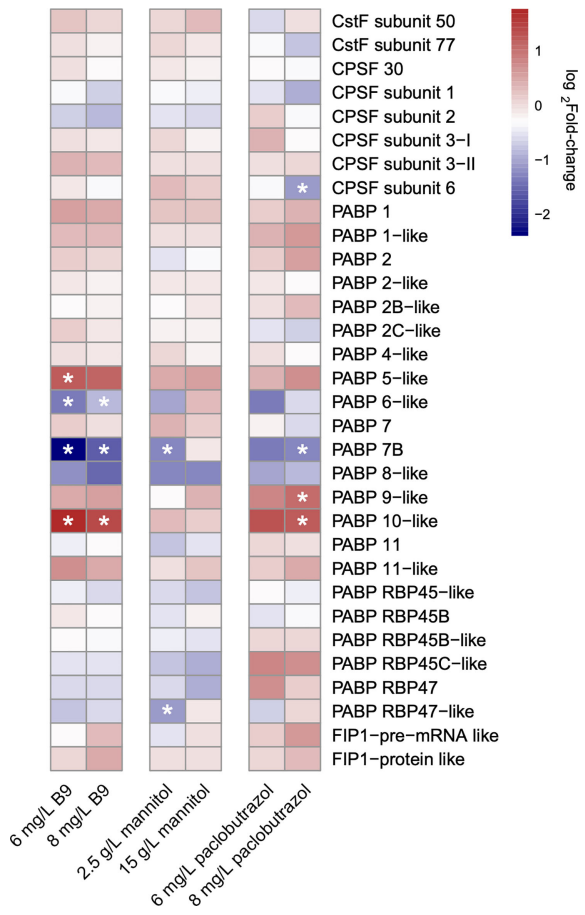


FIGURE 3
Heatmap illustrating the log2fold-change in the expression level of polyadenylation factor subunits under different treatments. * indicates the subunits of significantly differential expression under the treatment condition compared to CK.

572 out of the 582 and 2,553 DAGs, which accounted for 4.1% and 22.4%, were also differentially expressed under the treatment of 6 and 8 mg/L paclobutrazol, respectively (Figures 4A, D). The DE-specific genes induced by 6 mg/L paclobutrazol were highly represented in cell proliferation, auxin biosynthesis, and brassinosteroid (BR) response (Figure 4B), while the DA-specific genes were predominantly involved in growth regulation, pectinesterase activity, and responses to salicylic acid and strigolactone (Figure 4C). When exposed to 8 mg/L paclobutrazol, genes related to superoxide dismutase activity and root hair elongation were likely to be exclusively differentially expressed compared to CK (Figure 4E), whereas those participating in cellulose biosynthesis, oxidative stress regulation, and responses to salicylic acid and karrikin showed different APA usages (Figure 4F).

A similar pattern was also observed under the treatment of B9 and mannitol: only 4.2%–14.2% of DAGs were overlapped with DEGs (Supplementary Figures 2A, E and 3A, E). When treated with 8 mg/L B9, for one example, the DA-specific genes were significantly enriched in the processes of root development, maintenance of shoot apical meristem identity, and auxin transport (Supplementary Figure 2F), while the DE-specific genes were overrepresented in cell wall catabolism and oxidative stress responses (Supplementary Figure 2G). With regard to the application of 2.5 g/L mannitol, the DA-specific genes were highly represented in cellular response to osmotic stress, callose deposition in cell wall, and leaf senescence (Supplementary Figure 3B), and the GO terms related to cell proliferation, growth, and development were found to be enriched for the genes specifically regulated by differential expression (Supplementary Figure 3C). Comparatively, when treated with 15 g/L mannitol, the light-mediated leaf development, leaf senescence, and photoperiodism were mainly regulated by APA events (Supplementary Figure 3F), whereas the genes involved in auxin metabolism, cell development-related programmed cell death, cell wall thickening, secondary shoot formation, superoxide radical removal, and L-ascorbic acid transmembrane transport were largely under the control of different expression (Supplementary Figure 3G). Taken together, these results suggested that, in many scenarios, APA and gene transcription regulate different aspects of the growth and development of pomegranate seedlings and together contribute to the PGR-induced dwarfed stature.

Changes in 3′ UTR length affect microRNA target sites

APA events were also found to substantially modulate gene expression at post-transcriptional and translational levels. Compared to CK, DAGs in the PGR-treated seedlings displayed a global preference for using distal poly(A) sites (Figures 5A, B and Supplementary Figures 4A, 5A). For example, under the treatment of 8 mg/L paclobutrazol, 2,360 DAGs exhibited a higher abundance of the isoforms with longer 3′ UTRs, while only 305 genes used more proximal poly(A) sites (Figure 5B). These lengthened 3′ UTRs were supposed to host more miRNA target sites, which can further modulate mRNA abundance by influencing their stability. Consistently, 65.99%–74.07% of the isoforms using longer 3′ UTRs were inferred to consist of extra miRNA target sites, compared to those using shorter ones, under all the treatment scenarios (Figures 5A, B and Supplementary Figures 4B, 5B). For up to 61 isoforms, more than 10 putative miRNA targets were under the impact of the changes in 3′ UTR length (Figure 5C and Supplementary Figures 4C, 5C). Most of these miRNAs, both constitutive (existing in isoforms with both short and long 3′ UTRs) and lengthened 3′ UTR-specific miRNAs, were identified to function in cleavage of the corresponding mRNA, while 9.00%–12.73% of the miRNAs specific to the extended 3′ UTRs were supposed to inhibit mRNA translation (Figure 5D and Supplementary Figures 4D, 5D). More important, up to 259 DAGs with lengthened 3′ UTRs targeted by miRNAs were significantly downregulated under the treatment of PGRs (Supplementary Figures 4E, 5E, 6). When treated with 8 mg/L paclobutrazol, the miRNA-mediated downregulated genes were overrepresented in the GO terms of developmental growth, lateral root branching, maintenance of shoot apical meristem identity, and cellular response to strigolactone, indicative of their important roles in regulating the growth and development of pomegranate seedlings (Figure 5E).

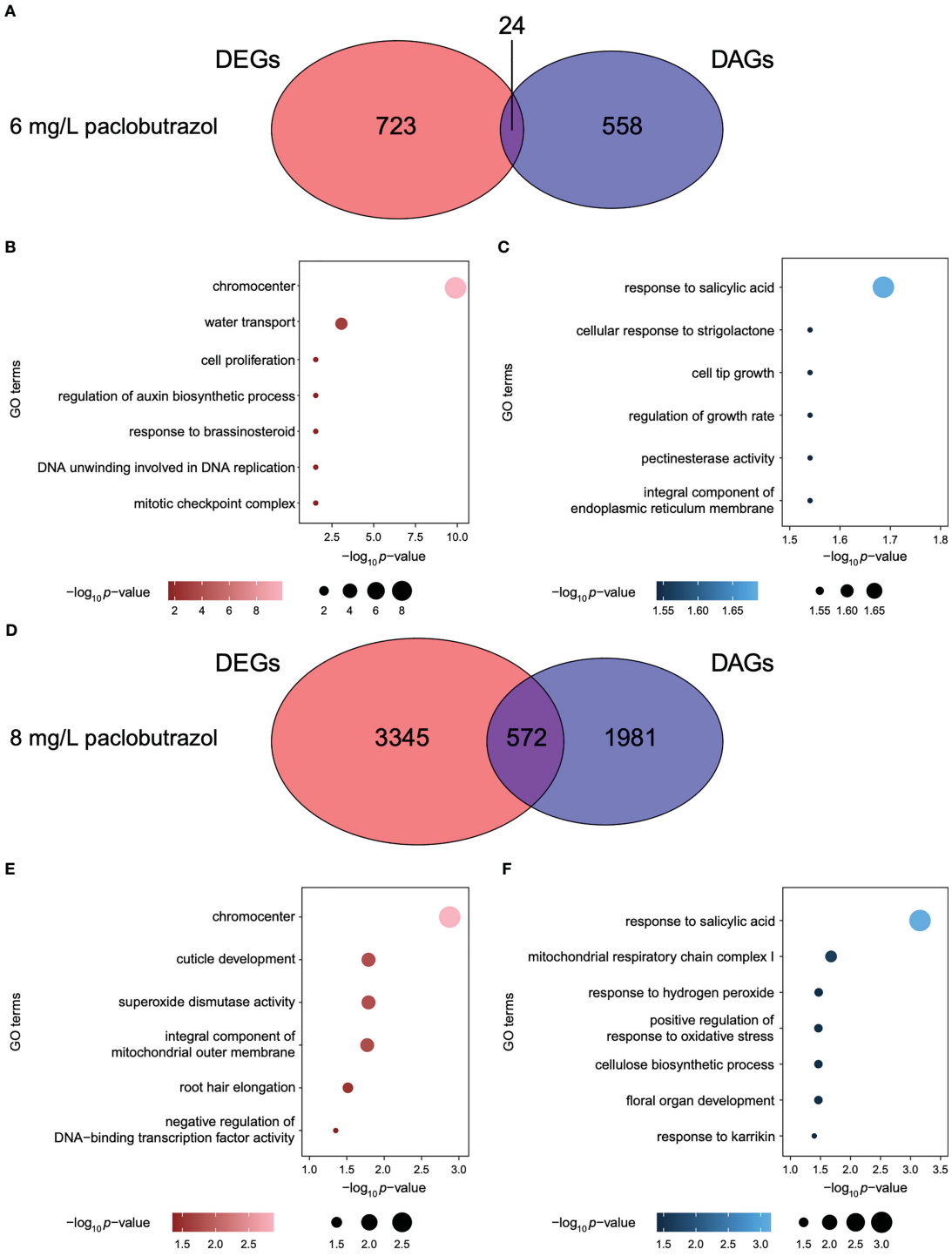


FIGURE 4
Comparison between DAGs and DEGs under the treatments of paclobutrazol. **(A)** Venn diagram illustrating the overlap between DAGs and DEGs when treated with 6 mg/L paclobutrazol. **(B, C)** GO terms enriched for the DA-specific **(B)** and DE-specific genes **(C)** under treatment of 6 mg/L paclobutrazol. **(D)** Venn diagram illustrating the overlap between DAGs and DEGs when treated with 8 mg/L paclobutrazol. **(E, F)** GO terms enriched for the DA-specific **(E)** and DE-specific genes **(F)** under the treatment of 8 mg/L paclobutrazol.

Discussion

In the current study, we explored the PGR-induced APA dynamics using the RNA-seq data from our previous study and showed that all the PGR treatments, even at low concentrations, provoked genome-wide alterations in the usages of poly(A) sites (Figure 1A). These changes were found to substantially influence how pomegranate seedlings grew and developed. For example, the poly(A) site usages of the genes involved in auxin transport and growth regulation were significantly altered after the treatments

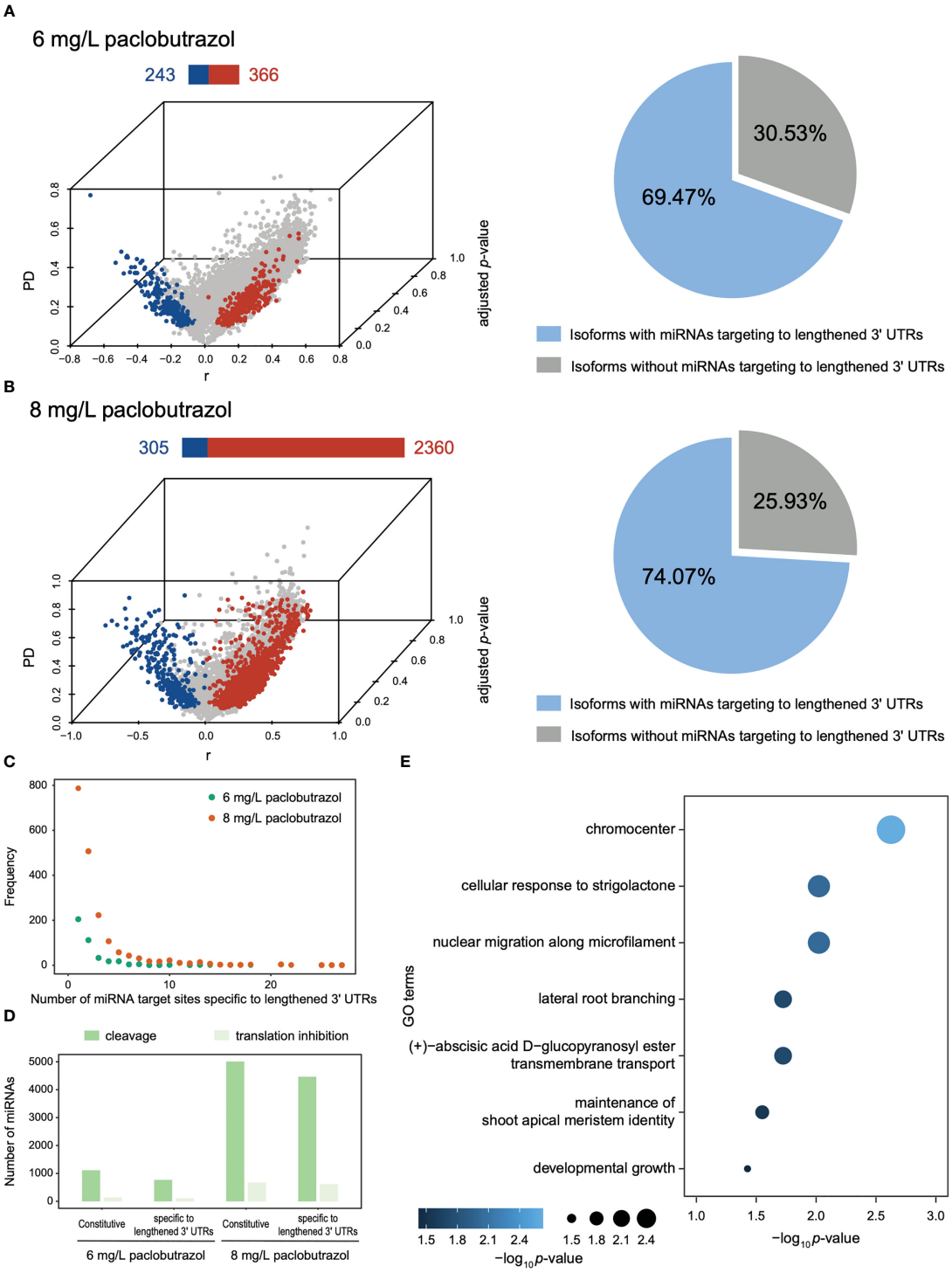


FIGURE 5
Overview of DAGs with lengthened/shortened 3′ UTR and the putative miRNAs targeting to the lengthened 3′ UTRs under the treatment of paclobutrazol. **(A, B)** Left panels: 3D volcano plot illustrating the DAGs displaying 3′ UTR lengthening (red) and shortening (blue) when treated with 6 **(A)** and 8 mg/L paclobutrazol **(B)**. The bar above each volcano plot shows the number of DAGs using more longer (red) and shorter 3′ UTRs (blue). Right panels: Pie charts showing the proportion of isoforms with (blue) or without (gray) putative miRNAs targeting the lengthened area of 3′ UTRs under 6 mg/L **(A)** and 8 mg/L paclobutrazol treatment **(B)**. **(C)** Frequency distribution illustrating the number of miRNA target sites identified specifically in lengthened 3′ UTRs across genes, when treated with 6 mg/L (green) and 8 mg/L paclobutrazol (orange), respectively. **(D)** Numbers of constitutive (existing in isoforms with both short and long 3′ UTRs) or lengthened 3′ UTR-specific miRNAs that were predicted with putative mRNA cleavage (green) and translation inhibition potentials (light green) under two paclobutrazol treatment scenarios, respectively. **(E)** Representative GO terms enriched for the DAGs that were with miRNAs targeting the lengthened 3′ UTR area and significantly downregulated in response to 8 mg/L paclobutrazol treatment.

(Figures 1B–D). In particular, one DAG encodes protein PIN-LIKES, which functions as an efflux carrier that mediates the unidirectional auxin flow (Polar auxin transport, PAT) among plant tissues (Křeček et al., 2009). The gene encoding mitogen-activated protein kinase kinase 2 (MKK2) also displayed significantly different APA profiles under the treatment. MKK2, together with mitogen-activated protein kinase 10 (MPK10), forms a module of mitogen-activated protein kinase (MAPK) signaling pathways that serves as a key regulator for PAT in plants (Jagodzik et al., 2018). These alterations in auxin transport may make contributions to the repressed growth and development in pomegranate. Correspondingly, the APA usages of the genes involved in shoot apical meristem identity maintenance and leaf development were also changed in response to PGR treatments (Figures 1, 2B). One of such genes is the calpain-type cysteine protease encoding gene *DEK1*. Studies in *Physcomitrella patens* highlighted the important function of *DEK1* in controlling the cell fate transition from 2D to 3D growth, where *DEK1* knockout in *P. patens* led to aberrant cell divisions and developmental arrest in buds (Demko et al., 2014; Johansen et al., 2016). Together, these widespread alterations in APA profiles under PGR treatments indicate the substantial significance of post-transcriptional mechanisms in modulating the dwarfed stature of pomegranate.

The APA dynamics display ample specificities among the treatments of different types/concentrations of PGRs (Figures 1, 2), which correspond to their distinct nature. The DAGs induced by 8 mg/L paclobutrazol were particularly associated with the response for karrikins, a type of plant growth regulator that controls plant development (Wang et al., 2020), while the genes with significant APA changes under 8 mg/L B9 treatment were overrepresented in phototropism and xanthophyll biosynthesis (Figure 2B). It is consistent with our previous observations from the transcriptome data that genes responsive to strigolactones, a type of plant signaling compound with similar biochemical properties and physiological activities to karrikins, were specifically downregulated when exposed to 8 mg/L paclobutrazol, whereas those involved in photosynthesis and photosystem II assembly/repair were suppressed by the application of B9 (Qian et al., 2022). Compared to B9 and paclobutrazol, mannitol treatments at both concentrations were found to elicit cellular response to osmotic stress (Figure 1C), corresponding to its specific mechanism that mannitol represses plant growth and development by increasing ambient osmatic pressure and causing drought stress to plants (Bhat and Chandel, 1993); thus, antioxidant reactions were activated to alleviate the oxidative damage. These results indicated that different regulatory mechanisms underlying the pomegranate dwarfing elicited by different PGRs were employed.

In the current study, we found that, in most scenarios, APA and transcriptional regulations are not synchronized and modulate PGR-induced growth repression *via* different routes, as manifested by both the little overlap between DAGs and DEGs and the differences in the pathways modulated by DAGs and DEGs (Figures 4A, D and Supplementary Figures 2A, E and 3A, E). The transcriptome data revealed that paclobutrazol obviously downregulated the genes related to the tryptophan-independent auxin biosynthetic process (Qian et al., 2022). Comparatively, APA events predominantly affected the polar transport of auxin among tissues. Similarly, when treated with 2.5 g/L mannitol, the cell wall modification process was modulated by the changes in both gene transcription and APA, although in distinct ways (Supplementary Figure 3). In particular, the gene encoding endoglucanase 8 (CEL1), which is a type of cellulose-hydrolyzing enzyme that regulates the cell wall relaxation associated with cell growth and expansion, was significantly downregulated (Tsabary et al., 2003; Shani et al., 2006). Consequently, the suppression of CEL1 would disrupt the differentiations of the plant vascular system and lead to shorter roots and shoots (Tsabary et al., 2003). In contrast, UTP-glucose-1-phosphate uridylyltransferase (also referred to as UDP-glucose pyrophosphorylase, UGPase), which supplies UDP-glucose substrate for the formation of secondary cell wall in plants, displayed substantial poly(A) usage variations. Moreover, Payyavula et al. (2014) showed that the maintenance of UGPase's function was important for the normal growth of *Populus deltoides*. These results suggested that both APA and gene transcription make key contributions to the intricate regulatory network underlying the dwarfing stature of pomegranate.

Despite the independent function of APA in regulation, APA is able to modulate transcriptome *via* influencing the presence or absence of regulatory elements located in 3' UTRs. Here, the DAGs induced by PGR treatments, especially at high concentrations, displayed a global preference for 3' UTR lengthening (Figures 5A, B left panel and Supplementary Figures 4A, 5A), which anchor more miRNA target sites than the corresponding shorter 3' UTRs (Figures 5A, B right panel and Supplementary Figures 4B, 5B). The majority of these extra "burdens" were supposed to suppress gene expression by causing the cleavage or destabilization of mRNA (Figure 5D and Supplementary Figures 4D, 5D). In particular, 259 DAGs with miRNAs specifically targeted in the lengthened 3' UTRs were significantly downregulated in response to the 8 mg/L paclobutrazol treatment (Supplementary Figure 6). These genes were found to participate in developmental growth, lateral root branching, and maintenance of shoot apical meristem (Figure 5E). Of them, the gene encoding E3 ubiquitin-protein ligase (KEG) is known by its negative regulatory activity of abscisic acid (ABA) signaling, and the suppression of its expression has been shown to retard the growth of *A. thaliana* (Stone et al., 2006). As another example, the expression of the gene that encodes ammonium transporter 1 member 1 (AMT1;1) was also supposed to be suppressed. AMT1;1 plays an important role in ammonium uptake from soil solution by roots and the subsequent root-to-shoot transport of ammonium; thus, the inhibition of AMT1;1 would lead to nitrogen deficiency and growth defect in pomegranate seedlings (Mayer and Ludewig, 2006). Together, these results highlight the role of APA events in fine-tuning gene expression in response to PGR treatments, which makes a key contribution to the retarded growth and development in the dwarfed pomegranate seedlings.

Conclusion

In this study, we, for the first time, identified and characterized the APA dynamics underlying PGR-elicited dwarfing in pomegranate. Our findings highlight the biological importance of post-transcriptional mechanisms in modulating pomegranate

growth and development, which adds a new dimension to the genetic basis of the agronomic trait of pomegranate. However, since our study is mainly based on the prediction from RNA-seq, we might be lacking in power to capture all of the signals and miss some of the true APA events. Thus, in the future, a more comprehensive investigation on the poly(A) usage alterations in pomegranate is essential using efficient technology to measure 3' UTR dynamics, such as Poly(A) tag sequencing (PAT-seq).

Data availability statement

The datasets we used in this study can be found in online repositories National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) with accession number of GSE195722.

Author contributions

YY and JQ designed the study. XX, MF, and YY collected the data and performed the bioinformatic analyses. XX, YL, XC, JW, JQ, and YY wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 32201420 to YY), the Key Research and Development Projects of Anhui province (No. 202204c06020062 to JQ), and the Production–Education–Research project (AKZY20 22110 to JQ).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1189456/full#supplementary-material

References

Asrey, R., Kumar, K., Sharma, R., and Meena, N. K. (2020). Fruit bagging and bag color affects physico-chemical, nutraceutical quality and consumer acceptability of pomegranate (Punica granatum l.) arils. *J. Food Sci. Technol.* 57, 1469–1476. doi: 10.1007/s13197-019-04182-x

Bhat, S., and Chandel, K. (1993). *In vitro* Conservation of musa germplasm: effects of mannitol and temperature on growth and storage. *J. Hortic. Sci.* 68 (6), 841–846. doi: 10.1080/00221589.1993.11516422

Bourekoua, H., Różyło, R., Gawlik-Dziki, U., Benatallah, L., Zidoune, M. N., and Dziki, D. (2018). Pomegranate seed powder as a functional component of gluten-free bread (Physical, sensorial and antioxidant evaluation). *Int. J. Food Sci. Technol.* 53 (8), 1906–1913. doi: 10.1111/ijfs.13777

Colgan, D. F., and Manley, J. L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* 11 (21), 2755–2766. doi: 10.1101/gad.11.21.2755

de Lorenzo, L., Sorenson, R., Bailey-Serres, J., and Hunt, A. G. (2017). Noncanonical alternative polyadenylation contributes to gene regulation in response to hypoxia. *Plant Cell* 29 (6), 1262–1277. doi: 10.1105/tpc.16.00746

Demko, V., Perroud, P.-F., Johansen, W., Delwiche, C. F., Cooper, E. D., Remme, P., et al. (2014). Genetic analysis of DEFECTIVE KERNEL1 loop function in three-dimensional body patterning in physcomitrella patens. *Plant Physiol.* 166 (2), 903–919. doi: 10.1104/pp.114.243758

Di Giammartino, D. C., Nishida, K., and Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Mol. Cell* 43 (6), 853–866. doi: 10.1016/j.molcel.2011.08.017

Elkon, R., Ugalde, A. P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* 14 (7), 496–506. doi: 10.1038/nrg3482

Foster, T. M., McAtee, P. A., Waite, C. N., Boldingh, H. L., and McGhie, T. K. (2017). Apple dwarfing rootstocks exhibit an imbalance in carbohydrate allocation and reduced cell growth and metabolism. *Hortic. Res.* 4, 17009. doi: 10.1038/hortres.2017.9

Hunt, A. G., Xing, D., and Li, Q. Q. (2012). Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. *BMC Genomics* 13, 641. doi: 10.1186/1471-2164-13-641

Jagodzik, P., Tajdel-Zielinska, M., Ciesla, A., Marczak, M., and Ludwikow, A. (2018). Mitogen-activated protein kinase cascades in plant hormone signaling. *Front. Plant Sci.* 9, 1387. doi: 10.3389/fpls.2018.01387

Johansen, W., Ako, A. E., Demko, V., Perroud, P.-F., Rensing, S. A., Mekhlif, A. K., et al. (2016). The DEK1 calpain linker functions in three-dimensional body patterning in physcomitrella patens. *Plant Physiol.* 172 (2), 1089–1104. doi: 10.1104/pp.16.00925

Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* 37 (8), 907–915. doi: 10.1038/s41587-019-0201-4

Křeček, P., Skůpa, P., Libus, J., Naramoto, S., Tejos, R., Friml, J., et al. (2009). The PIN-FORMED (PIN) protein family of auxin transporters. *Genome Biol.* 10 (12), 1–11. doi: 10.1186/gb-2009-10-12-249

Luo, X., Li, H., Wu, Z., Yao, W., Zhao, P., Cao, D., et al. (2020). The pomegranate (Punica granatum l.) draft genome dissects genetic divergence between soft-and hard-seeded cultivars. *Plant Biotechnol. J.* 18 (4), 955–968. doi: 10.1111/pbi.13260

Mayer, M., and Ludewig, U. (2006). Role of AMT1;1 in NH4+ acquisition in arabidopsis thaliana. *Plant Biol.* 8 (4), 522–528. doi: 10.1055/s-2006-923877

Payyavula, R. S., Tschaplinski, T. J., Jawdy, S. S., Sykes, R. W., Tuskan, G. A., and Kalluri, U. C. (2014). Metabolic profiling reveals altered sugar and secondary metabolism in response to UGPase overexpression in populus. *BMC Plant Biol.* 14, 1–14. doi: 10.1186/s12870-014-0265-8

Qian, J., Wang, N., Ren, W., Zhang, R., Hong, X., Chen, L., et al. (2022). Molecular dissection unveiling dwarfing effects of plant growth retardants on pomegranate. *Front. Plant Sci.* 13, 866193. doi: 10.3389/fpls.2022.866193

Qian, J., Zhang, X., Yan, Y., Wang, N., Ge, W., Zhou, Q., et al. (2020). Unravelling the molecular mechanisms of abscisic acid-mediated drought-stress alleviation in pomegranate (Punica granatum l.). *Plant Physiol. Biochem.* 157, 211–218. doi: 10.1016/j.plaphy.2020.10.020

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26 (6), 841–842. doi: 10.1093/bioinformatics/btq033

Seleznyova, A. N., Tustin, D. S., and Thorp, T. G. (2008). Apple dwarfing rootstocks and interstocks affect the type of growth units produced during the annual growth cycle: precocious transition to flowering affects the composition and vigour of annual shoots. *Ann. Bot.* 101 (5), 679–687. doi: 10.1093/aob/mcn007

Shahamirian, M., Eskandari, M. H., Niakousari, M., Esteghlal, S., Hashemi Gahruie, H., and Mousavi Khaneghah, A. (2019). Incorporation of pomegranate rind powder extract and pomegranate juice into frozen burgers: oxidative stability, sensorial and microbiological characteristics. *J. Food Sci. Technol.* 56, 1174–1183. doi: 10.1007/s13197-019-03580-5

Shani, Z., Dekel, M., Roiz, L., Horowitz, M., Kolosovski, N., Lapidot, S., et al. (2006). Expression of endo-1, 4-β-glucanase (cel 1) in arabidopsis thaliana is associated with plant growth, xylem development and cell wall thickening. *Plant Cell Rep.* 25, 1067–1074. doi: 10.1007/s00299-006-0167-9

Shen, Y., Ji, G., Haas, B. J., Wu, X., Zheng, J., Reese, G. J., et al. (2008). Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation. *Nucleic Acids Res.* 36 (9), 3150–3161. doi: 10.1093/nar/gkn158

Singh, A., and Roychoudhury, A. (2021). Gene regulation at transcriptional and post-transcriptional levels to combat salt stress in plants. *Physiol. Plant* 173 (4), 1556–1572. doi: 10.1111/ppl.13502

Stone, S. L., Williams, L. A., Farmer, L. M., Vierstra, R. D., and Callis, J. (2006). KEEP ON GOING, a RING E3 ligase essential for arabidopsis growth and development, is involved in abscisic acid signaling. *Plant Cell* 18 (12), 3415–3428. doi: 10.1105/tpc.106.046532

Sun, Y., Fu, Y., Li, Y., and Xu, A. (2012). Genome-wide alternative polyadenylation in animals: insights from high-throughput technologies. *J. Mol. Cell Biol.* 4 (6), 352–361. doi: 10.1093/jmcb/mjs041

Tian, B., and Manley, J. L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* 18 (1), 18–30. doi: 10.1038/nrm.2016.116

Tsabary, G., Shani, Z., Roiz, L., Levy, I., Riov, J., and Shoseyov, O. (2003). Abnormalwrinkled'cell walls and retarded development of transgenic arabidopsis thaliana plants expressing endo-1, 4-[beta]-glucanase (cell) antisense. *Plant Mol. Biol.* 51 (2), 213. doi: 10.1023/A:1021162321527

Turrini, F., Boggia, R., Donno, D., Parodi, B., Beccaro, G., Baldassari, S., et al. (2020). From pomegranate marcs to a potential bioactive ingredient: a recycling proposal for pomegranate-squeezed marcs. *Eur. Food Res. Technol.* 246, 273–285. doi: 10.1007/s00217-019-03339-4

Wang, T., Liu, L., Wang, X., Liang, L., Yue, J., and Li, L. (2018). Comparative analyses of anatomical structure, phytohormone levels, and gene expression profiles reveal potential dwarfing mechanisms in shengyin bamboo (Phyllostachys edulis f. tubaeformis). *Int. J. Mol. Sci.* 19 (6), 1697. doi: 10.3390/ijms19061697

Wang, L., Xu, Q., Yu, H., Ma, H., Li, X., Yang, J., et al. (2020). Strigolactone and karrikin signaling pathways elicit ubiquitination and proteolysis of SMXL2 to regulate hypocotyl elongation in arabidopsis. *Plant Cell* 32 (7), 2251–2270. doi: 10.1105/tpc.20.00140

Wang, T., Ye, W., Zhang, J., Li, H., Zeng, W., Zhu, S., et al. (2023). Alternative 3′-untranslated regions regulate high-salt tolerance of spartina alterniflora. *Plant Physiol.* 191 (4), 2570–2587. doi: 10.1093/plphys/kiad030

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q. Q., et al. (2011). Genome-wide landscape of polyadenylation in arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 108 (30), 12533–12538. doi: 10.1073/pnas.1019732108

Ye, C., Long, Y., Ji, G., Li, Q. Q., and Wu, X. (2018). APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* 34 (11), 1841–1849. doi: 10.1093/bioinformatics/bty029

Ye, C., Zhou, Q., Wu, X., Ji, G., and Li, Q. Q. (2019). Genome-wide alternative polyadenylation dynamics in response to biotic and abiotic stresses in rice. *Ecotoxicol. Environ. Saf.* 183, 109485. doi: 10.1016/j.ecoenv.2019.109485

Yu, Z., Hong, L., and Li, Q. Q. (2022). Signatures of mRNA alternative polyadenylation in arabidopsis leaf development. *Front. Genet.* 13, 863253. doi: 10.3389/fgene.2022.863253

Zhou, Q., Fu, H., Yang, D., Ye, C., Zhu, S., Lin, J., et al. (2019). Differential alternative polyadenylation contributes to the developmental divergence between two rice subspecies, japonica and indica. *Plant J.* 98 (2), 260–276. doi: 10.1111/tpj.14209

Zhou, Y., and Underhill, S. J. R. (2021). Differential transcription pathways associated with rootstock-induced dwarfing in breadfruit (Artocarpus altilis) scions. *BMC Plant Biol.* 21 (1), 261. doi: 10.1186/s12870-021-03013-6

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 29 May 2023
DOI 10.3389/fpls.2023.1179357

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Hengyou Zhang,
Chinese Academy of Sciences (CAS), China
Sang He,
Chinese Academy of Agricultural Sciences, China

*CORRESPONDENCE
Haizheng Xiong
hxx007@uark.edu
Jinshe Wang
wjs33314@126.com
Ainong Shi
ashi@uark.edu

RECEIVED 04 March 2023
ACCEPTED 25 April 2023
PUBLISHED 29 May 2023

CITATION
Xiong H, Chen Y, Pan Y-B, Wang J, Lu W and Shi A (2023) A genome-wide association study and genomic prediction for *Phakopsora pachyrhizi* resistance in soybean.
Front. Plant Sci. 14:1179357.
doi: 10.3389/fpls.2023.1179357

A genome-wide association study and genomic prediction for *Phakopsora pachyrhizi* resistance in soybean

Haizheng Xiong[1*], Yilin Chen[1], Yong-Bao Pan[2], Jinshe Wang[3*], Weiguo Lu[3] and Ainong Shi[1*]

[1]Department of Horticulture, University of Arkansas, Fayetteville, AR, United States, [2]Sugarcane Research Unit, Untied State Department of Agriculture – Agriculture Research Service (USDA-ARS), Houma, LA, United States, [3]Henan Academy of Crops Molecular Breeding, National Centre for Plant Breeding, Zhengzhou, China

Soybean brown rust (SBR), caused by *Phakopsora pachyrhizi*, is a devastating fungal disease that threatens global soybean production. This study conducted a genome-wide association study (GWAS) with seven models on a panel of 3,082 soybean accessions to identify the markers associated with SBR resistance by 30,314 high quality single nucleotide polymorphism (SNPs). Then five genomic selection (GS) models, including Ridge regression best linear unbiased predictor (rrBLUP), Genomic best linear unbiased predictor (gBLUP), Bayesian least absolute shrinkage and selection operator (Bayesian LASSO), Random Forest (RF), and Support vector machines (SVM), were used to predict breeding values of SBR resistance using whole genome SNP sets and GWAS-based marker sets. Four SNPs, namely Gm18_57,223,391 (LOD = 2.69), Gm16_29,491,946 (LOD = 3.86), Gm06_45,035,185 (LOD = 4.74), and Gm18_51,994,200 (LOD = 3.60), were located near the reported P. pachyrhizi R genes, Rpp1, Rpp2, Rpp3, and Rpp4, respectively. Other significant SNPs, including Gm02_7,235,181 (LOD = 7.91), Gm02_7234594 (LOD = 7.61), Gm03_38,913,029 (LOD = 6.85), Gm04_46,003,059 (LOD = 6.03), Gm09_1,951,644 (LOD = 10.07), Gm10_39,142,024 (LOD = 7.12), Gm12_28,136,735 (LOD = 7.03), Gm13_16,350,701(LOD = 5.63), Gm14_6,185,611 (LOD = 5.51), and Gm19_44,734,953 (LOD = 6.02), were associated with abundant disease resistance genes, such as *Glyma.02G084100*, *Glyma.03G175300*, *Glyma.04g189500*, *Glyma.09G023800*, *Glyma.12G160400*, *Glyma.13G064500*, *Glyma.14g073300*, and *Glyma.19G190200*. The annotations of these genes included but not limited to: *LRR* class gene, *cytochrome 450*, cell wall structure, *RCC1, NAC, ABC* transporter, *F-box* domain, etc. The GWAS based markers showed more accuracies in genomic prediction than the whole genome SNPs, and Bayesian LASSO model was the ideal model in SBR resistance prediction with 44.5% ~ 60.4% accuracies. This study aids breeders in predicting selection accuracy of complex traits such as disease resistance and can shorten the soybean breeding cycle by the identified markers

KEYWORDS

GWAS, soybean, disease resistance, genomic prediction, *Phakopsora pachyrhizi*

Introduction

Soybean brown rust (SBR) is one of the most devastating fungal diseases of soybean (*Glycine max*) (Hartman et al., 2005). It first emerged around 1900 as a threat to soybean production in China and Japan and has since spread globally, in part due to human activities and meteorological phenomena (Hartman et al., 1991). The disease arrived in Africa and the Pacific Islands in the 1980s and 1990s and later reached the American continents in the 2000s (Miles et al., 2004). The risk of SBR attracted more attention with the disease outbreak in China in 1975 and in Brazil in 2001, that caused 10 billion US dollar losses in each country (Yorinori et al., 2005; Godoy et al., 2016). Comparing to the native American rust pathogen (*Phakopsora meibomiae*), the exotic one (*Phakopsora pachyrhizi*) was much more aggressive and caused an epidemic on soybean in South America and spread to North America (Pivonia and Yang, 2004).

Soybean plants are susceptible to SBR at any stage of growth and development and *Phakopsora pachyrhizi* can quickly spread over a long-range through wind-borne urediniospores (Isard et al., 2005). Therefore, it is important to develop control strategies for controlling SBR. Currently, the SBR can be managed by applying fungicides and employing specific cultivation practices (Levy, 2005). However, considering the high cost and the harm to non-target beneficial fungi, a more economic, safer, and environmental friendly solution is to raise varieties' own resistance by developing new resistance lines through breeding or engineering (Bromfield and Hartwig, 1980). In the past 30 years, the well-known *Rpp 1–7* genes were mapped to chromosome 3, 6, 16, 18, and 19 (Garcia et al., 2008; Pandey et al., 2011; Li et al., 2012; Kashiwa et al., 2020). However, *Rpp* genes were race-specific and provided resistance exclusively to specific *P. pachyrhizi* isolates. Currently, there is no resistant soybean genotype that can ward off all known *P. pachyrhizi* isolates (Childs et al., 2018a). In addition, *Rpp* gene-mediated resistance can be overcome swiftly in the field due to pathogen's adaptation and evolution to resistant host (Godoy and Meyer, 2020). Pyramiding three or more *Rpp* genes into one genotype to obtain broader and/or more durable resistance has been reported on other crops like wheat or barley, but traditional breeding is still time-consuming and may introduce unwanted traits (Childs et al., 2018a). Another promising strategy for sustainable and effective SBR resistance is to utilize alternative R gene combinations and dynamic turnover in the field (Childs et al., 2018a). However, the identity of these *Rpp* genes needs to be revealed (Gebremedhn et al., 2020). Under the current conditions, it is also impractical to rely only on several major genes or combinations of these genes to control the SBR disease in field production.

In addition to major genes, many recent molecular studies have revealed more disease-resistant pathways in soybeans (Childs et al., 2018b). The resistance usually occurs in the form of signals, transcription factors, NB-LRR, or secondary metabolites (Gebremedhn et al., 2020; Waheed et al., 2021). They usually improve not only the resistance to a particular pathogen but the overall resistance of the plant as well. In addition, many minor resistance/tolerance genes are widely distributed throughout the whole soybean genome and exhibit partial defense response (PDR) to SBR (Langenbach et al., 2016). PDR is characterized by low infection frequency, long-lasting latency, small lesions, and reduced spore production per uredinium (Langenbach et al., 2016). At the molecular level, their specific functions are sometimes very similar or overlapping to the context components; however, they are more complex and obscure (Langenbach et al., 2016). Screening for or silencing susceptibility is another strategy that can provide durable disease resistance in breeding, because of susceptible (S) gene function either as susceptibility factors or suppressors of plant defense, thus potential targets of fungal effectors (De Wit, 1992). For example, absence of the S gene *Mlo* in barley results in an incompatibility interaction with *Blumeria graminis hordei* that resembles nuclear hormone receptors (Büschges et al., 1997; Lucas, 2020). However, the identification and mapping of S gene are more difficult than those of major R genes by linkage mapping, and only one [Cys(2)His(2) zinc finger TF palmate-like pentafoliata1, PALM1] would classify as a S gene so far (Uppalapati et al., 2012).

Molecular marker-assisted selection (MAS) has been applied in soybean breeding to accelerate the development of disease-resistant varieties, and the GWAS is of vital help to MAS (He et al., 2014). Comparing with linkage mapping, GWAS can not only find the major genes, but also has the incomparable ability to map and identify the minor and S genes. Moreover, since the mapping populations such as natural population and multi-parent advanced generation inter-cross, contain more diversity, the markers developed have more universal applicability (Visscher et al., 2012). So far, only one SBR-related GWAS has been reported by Chang et al. (2016), who used GWAS to discover five SBR-related loci from USDA germplasm. Genomic selection (GS) has gained popularity in recent years in modern and large-scale crop breeding programs. GS can predict the breeding value of an individual plant based on its genotype to estimate the field performance of the plant, whereas MAS relies on the detection of a few QTLs using a simple linear model. Therefore, molecular breeding would shift from marker-assisted selection to genomic selection, as the genetic architecture of resistance changes from a single major R gene to multiple minor diffusion gene architectures (Poland and Rutkoski, 2016). Additionally, GS has been reported to be a useful tool in soybean breeding to predict a wide range of traits, including both agronomic and quality traits (Lorenz et al., 2011). However, no research has been done with respect to investigating GS accuracy for SBR resistance/tolerance.

The objectives of this study were to identify SBR resistance-associated SNP markers and to characterize the ability of genomic prediction in order to use SNP markers in selecting soybean breeding lines highly resistant to SBR.

Materials and methods

Plant materials and phenotyping

SBR disease scores and phenotyping data of 3,082 soybean accessions (Table S1) were downloaded from the USDA GRIN

website (https://npgsweb.ars-grin.gov/gringlobal/method?id=492634) (Miles et al., 2006). Based on the website, a greenhouse study was initiated. Soybean plants of 3,082 accessions were spray-inoculated between the first and second trifoliate stage with a mixture of urediniospores (60,000 spores per ml) from four *Phakopsora pachyrhizi* isolates, incubated overnight in a dew chamber at 22–25°C, and placed in a greenhouse at 20–25°C for 14 days. Disease severity was evaluated on the first trifoliate leaves for most accessions; however, the unifoliate leaves were evaluated for a few accessions due to slow germination (Miles et al., 2006). Based on the symptom and lesion development, a disease severity scale of 1 to 5 was used, where 1 = no visible symptom, 2 = light infection: only a few small (less than 1 cm) rust lesion present on the leaves, 3 = light to moderate infection: moderately sized (1–2 cm) rust lesion present on a limited number of leaves, 4 = moderate to severe infection: large (greater than 2 cm) rust lesion present on a significant number of leaves, and 5 = severe infection: nearly all leaves are covered in large rust lesion, and the disease is causing a significant damage to the plant growth (Walker et al., 2011).

Genotyping

The Soy50K SNP Infinium Chips (Song et al., 2013) and a total of 42,292 SNPs across 3,082 soybean accessions were downloaded from the Soybase at https://www.soybase.org/snps/download.php. SNPs with >10% missing data, >8% heterozygous genotypes, and <10% minor allele frequencies (MAF) were removed, and 30,314 SNPs were included in the GWAS study.

Population structure and genetic diversity

LEA is an R package for population structure and genomic signature analysis of local adaptation. The inference algorithms used by R are based on a fast version of structure available from the R package LEA (Frichot and François, 2015). The structure analysis identifies K clusters by measuring an optimum ΔK based on the SNP data provided. A preliminary analysis was performed in multiple runs by inputting successive values of K from 2 to 20. After an optimum K was determined, each soybean accession was assigned to a cluster (Q) based on the probability that the accession belonged to that cluster. The cut-off probability for the assignment to a cluster was 0.5. Based on the optimum K, a bar plot with "Sort by Q" was obtained to visualize the population structure among the 3,082 accessions. Phylogenetic relationships among the accessions was generated by TASSEL 5.2.13 and phylogenetic tree was drawn using R package: Phytologist and Phytools (Revell, 2012). During the drawing of the phylogeny trees, the population structure and the cluster information were imported for the combined analysis of genetic diversity. For sub-tree of each Q (cluster), the shape of "Node/Subtree Marker" and the "Branch Line" was drawn using the same color scheme of the STRUCTURE analysis.

Linkage disequilibrium analysis and SNP based haplotype blocks

TASSEL 5.0 (Bradbury et al., 2007) was used to calculate the linkage disequilibrium (LD) for all pairwise loci. Only SNPs with a minor allele frequency (MAF) greater than 0.10 and less than 10% missing data were included in the LD estimation process. Haplotype blocks (HAP) were estimated by Plink 2.0 (Purcell et al., 2007) within 200kb ($r^2 \approx 0.4$), and a minimum threshold value 0.05 for MAF.

Genome-wide association study

GWAS was performed using the Generalized Linear Model (GLM), Mixed Linear Model (MLM) (Jiang and Nguyen, 2021), Compressed Mixed Linear Model (CMLM), Multiple Loci Mixed Model (MLMM) (Wen et al., 2018), Settlement of MLM Under Progressively Exclusive Relationship (SUPER) (Wang et al., 2014), Fixed and Random Model Circulating Probability Unification (FarmCPU) (Liu et al., 2016), and Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK) (Wang et al., 2014) in R software GAPIT 3 (Genomic Association and Prediction Integrated Tool version 3) (Wang and Zhang, 2021; Lipka et al., 2012; https://zzlab.net/GAPIT/index.html; https://github.com/jiabowang/GAPIT3) by setting PCA = 6, with the Kinship for MLM, CMLM, MLMM, SUPER; and Pseudo QTNs for FarmCPU and BLINK.

SNP selection accuracy and selection efficiency

The accuracy and efficiency of SNP selection were computed to evaluate the significant SNP by the allele proportion in the population.

Selection accuracy (SA) = 100*[(Number of S or R genotypes with the favorable SNP allele)/(Number of R genotypes with the favorable SNP allele + Number of S genotypes with the favorable SNP allele)]/ΔE, where $\Delta E = E_1/E_2$, E_1 = Observed number of S or R genotypes/(Number of R genotypes + Number of S genotypes); E_2= Expected number of S or R genotypes/(Number of R genotypes + Number of S genotypes).

Selection efficiency (SE) =100*[(Number of S or R genotypes with the favorable SNP allele)/(Total number of S or R genotypes)]/ΔF, where $\Delta F = F_1/F_2$, F_1 = Observed allele frequency of SNP, and F_2 = Expected allele frequency of SNP. In this study we set the E_2 and F_2 as an ideal equilibrium value (50%).

Candidate gene prediction

Candidate genes were selected based on the peak significant SNP in each LD region located within 50 kb on either side of significant SNPs (Zhang et al., 2016), and furtherly by 0 kb (on the gene), 1 kb, 5 kb, 10 kb, 20 kb, 30 kb, and 50 kb, respectively. Candidate genes were

retrieved from the reference annotation of the soybean reference genome Wm82.a2.v1 from the SoyBase (http://www.soybase.org) and the Phytozome database (https://phytozome.jgi.doe.gov).

Genomic prediction

GP was conducted using seven SNP sets: All SNP set (30,314 SNPs) and six GWAS-derived SNP marker sets. The six GWAS-derived SNP marker sets consisted of those significant SNPs from highest LOD [–log(P-value)] to low LOD value (2.0) to make each set as 28, 100, 500, 1,000, 2,000, and 5,000 SNPs, respectively. Genomic estimated breeding value (GEBV) was computed using five statistical models: Ridge regression best linear unbiased predictor (rrBLUP) (Endelman, 2011), Genomic best linear unbiased predictor (gBLUP) (Zhang et al., 2007), Bayesian least absolute shrinkage and selection operator (Bayesian LASSO) (Heslot et al., 2012), Random Forest (RF) (Poland et al., 2012), and Support vector machines (SVM) (Ogutu et al., 2011) (Table S2).

A five-fold cross-validation was performed for each GP. The association panel was randomly divided into 5 disjoint subsets, 4 subsets were used as training set, and the remaining set was considered testing set. A total of 100 replications were conducted at each fold. Mean and standard errors corresponding to each fold were computed. Genomic prediction accuracy was obtained by computing the Pearson's correlation coefficient (r) between GEBV and the observed phenotypic value for the testing set (Shikha et al., 2017).

Results

Germplasm evaluation of *Phakopsora pachyrhizi*

Out of 3,082 soybean accessions evaluated for TAN lesion type, 71 (2.3%) were rated 1~2, 1,009 (32.7%) were rated 2.3~3, 1,746 (56.7%) were rated 3.1~4; and 256 (8.3%) were rated 4.2~5 in a rating scale of 1 to 5. Accessions with a mean severity of 2.7 or less (299, 9.5%) were considered resistant, while those with a mean severity of 4.0 or more (791, 25.6%) were considered susceptible. Accessions between the two categories were considered moderate. There was a large range in the distribution of each category (Figure 1). Majority of accessions displayed a disease severity rating of 3 or 4 being susceptible to SBR.

SNP profile

A total of 30,314 high quality SNPs were used to perform GWAS in the soybean accessions. Number of SNPs per chromosome ranged from 1,027 on chr20 to 1,898 on Chr16, with an average of 1,515.7 SNPs (Figure 2). The average distance between two SNPs per chromosome varied from 23.6 kb to 46.6 kb, with an average of 33.1 kb. The shortest average distance between SNPs was found on Chr18, whereas the longest one was on Chr20. Average MAF per chromosome ranged between 25.8% and 30.1%, with an average of 28.7% (Table S3). Percentage of heterozygous SNPs across all chromosomes were below 0.7%, and the percentage of missing SNPs per chromosome varied from 0.3% to 0.7%.

Population structure and LD haplotype

The structure analysis helped identify the most promising genetic variations to better understand the genetic basis of the trait. The population structure of the soybean accessions was analyzed using the R packages "LEA" and the peak of ΔK was observed at K = 6, indicating of the presence of six subpopulations or clusters (Figure 3A). A total of 337 (10.9%) accessions were assigned to subpopulation Q1; 306 (9.9%) assigned to Q2; 543 (17.6%) assigned to Q3; 534 (17.3%) assigned to Q4; 358 assigned to Q5; and 1,004 (32.5%) assigned to Q6 (Figure 3B). Phylogenetic analysis and population admixture map using R packages "Phytool" and "LEA" also revealed that the clustering of accessions was consistent with that inferred by structure K = 6 (Figure 3C). Additionally, there was a clear tendency of clustering by geographical areas. The controlling for population structure by taking geography into account is crucial for accurate GWAS results and for identifying true genetic associations with the trait

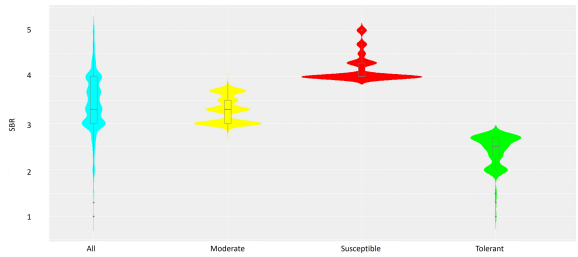


FIGURE 1
Combined violin-boxplots based on SBR ranking of the 3,082 soybean accessions, including Susceptible (red), Moderate (yellow) and Tolerant (green) groups.

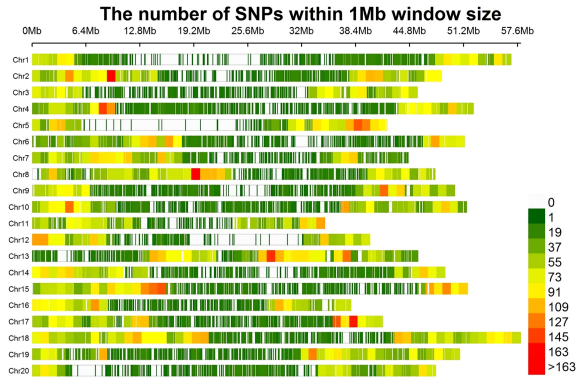


FIGURE 2
The distribution of 30,314 SNPs among the 20 chromosomes of soybean within 1 Mb size.

of interest. As Q6 was dominant in South and Central China and Southeast Asia, Q3 and Q4 were main populations in Northeast and Northwest Asia, and the population in Europe was dominated by Q2 and Q5 (Figure 3D and Table S1). Kinship matrix, based on 30,314 polymorphic SNPs for the studied genotypes, indicated that there was no clear clustering among the 3082 genotypes.

We examined the linkage disequilibrium (LD) decay patterns by 30,314 genome-wide SNPs. To visualize the LD decay patterns across distances, we plotted the LD decay curves by GAPIT 3 (Figure 4). The LD decay curves showed a clear distance-dependent pattern, with steeper decay curves at longer distances. Specifically, at a distance of 103 kb, the LD decayed with an R^2 value of 0.6, indicating a relatively strong LD correlation between nearby variants. At 216 kb, the LD decayed with an R^2 value of 0.4, indicating a moderate level of LD correlation between nearby variants. Finally, at 296 kb, the LD decayed with an R^2 value of 0.2, indicating a weak level of LD correlation between nearby variants (Figure S1). A total of 4,940 haplotype blocks were identified based on 30,314 SNPs. Number of blocks per chromosome varied from 170 on Chr11 to 357 on Chr18. Number of SNPs within each block varied from 2 to 67. Many haplotype blocks contained more than two significant SNP markers, for example, Gm01_47,462,126, Gm01_47,476,910, Gm01_47,481,216, Gm01_47,495,955, Gm01_47,503,665, Gm01_47,516,500, and Gm01_47,548,257 were in the same haplotype block on Chr1 (Table S4).

Genome-wide association study

The high convergence and consistency of the GWAS were observed among seven models. For example, the top six significant SNPs from the FarmCPU model including: Gm09_1,951,644 (10.06), Gm20_36,724,867 (6.54), Gm03_38,913,029 (6.10), Gm19_44,734,953 (5.7), Gm02_7,235,181 (5.18), and Gm04_47,132,429 (5.06) also had the high LOD values, which were at least 5.20, 2.67, 3.77, 3.59, 3.69, and 4.00 in other models. SNPs Gm04_45,884,688, Gm10_39,142,024, Gm14_2,492,139, Gm16_4,935,328, etc. were significant among all seven models (Figures 4, S2). A total of 100 SNPs were collected in this study by considering both model consistency and significance (Table S5). These SNPs were positioned at 47 haplotype blocks (Table S4). Then the top 28 SNPs with LOD > 5.50 were listed in Table 1 for future discussion. These 28 SNPs were located on 13 chromosomes (Chr. 2, 3, 4, 6, 8, 9, 10, 12, 13, 14, 16, 19, and 20), indicating their wide distribution and presence of genes that confer SBR resistance across the genome. Several SNPs were found in the same blocks, such as Gm02_7,235,181 and Gm02_7,234,594 in block 436; Gm09_1,944,730, Gm09_1,943,831, and Gm09_1,951,644 in block 1902; Gm10_5,573,877, Gm10_5,573,007, Gm10_5,559,592, Gm10_5,541,691, and Gm10_5,578,693 in block 2331; and Gm10_39,142,024 and Gm10_3,9147,121 in block 2215, which might be due to the gene clustering or pleiotropy.

Candidate genes of significant SNPs

Due to variations in LD decay across different regions, a conservative distance of 50 kb was set to select candidate genes as the region of the significant SNPs. There are four SNPs (loci) out of the top 100 associated markers, including Gm18_57,223,391,

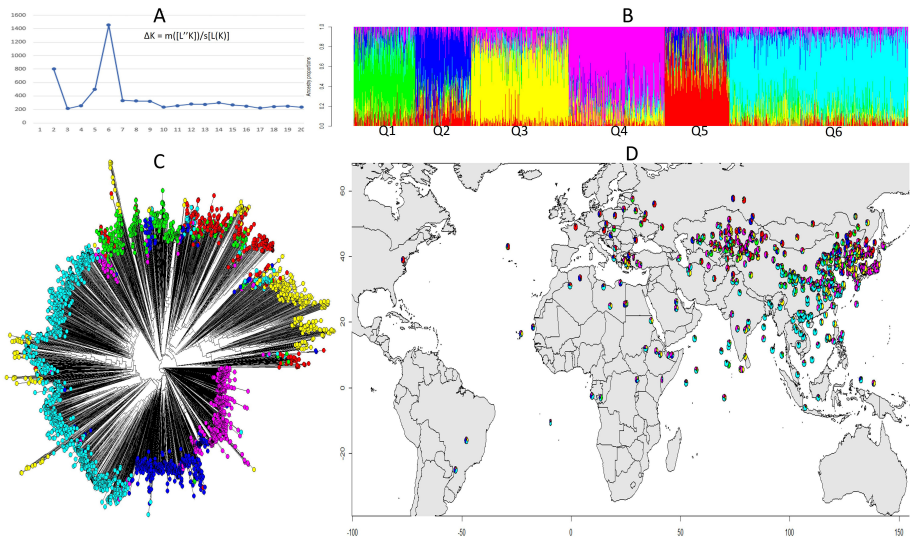


FIGURE 3
Structural and phylogenetic analysis of 3,082 soybean accessions based on 30,314 SNPs. **(A)** Delta K values for different numbers of populations assumed (K=20) in the STRUCTURE analysis. **(B)** Classification of soybean accessions in six groups (K=6) using STRUCTURE. The distribution of accessions to different populations is color coded, Q1 (green), Q2 (blue), Q3 (yellow), Q4 (pink), Q5 (red), Q6 (cyan). The x-axis shown the accessions of each subgroup, and the number on the y-axis shows the Q likelihood of accessions. **(C)** Phylogenetic analysis of the 3,082 soybean accessions with the corresponded labels used in **(B)**. **(D)** Geographical distribution of the soybean accessions by colored pie chart corresponding with the group proportion **(B)**.

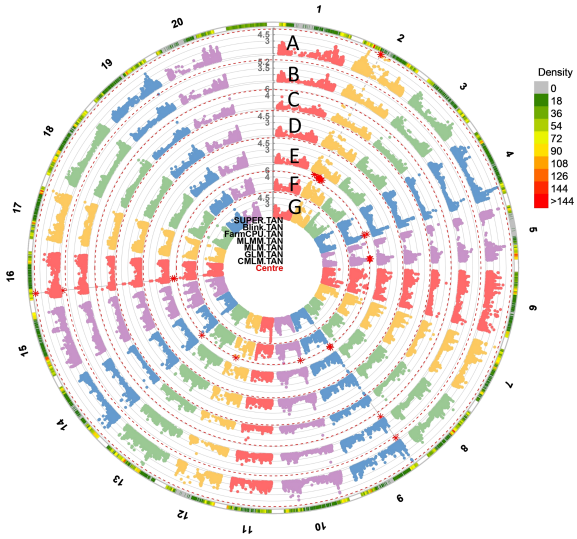


FIGURE 4
The circular Manhattan plots of seven GWAS models: **(A)** Settlement of MLM Under Progressively Exclusive Relationship (SUPER), **(B)** Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway (BLINK), **(C)** Fixed and Random Model Circulating Probability Unification (FarmCPU), **(D)** Multiple Loci Mixed Model (MLMM), **(E)** Mixed Linear Model (MLM), **(F)** Generalized Linear Models (GLM) and **(G)** Compressed Mixed Linear Model (CMLM) for SBR. The red asterisk points to the significant spots associated with SBR on 20 chromosomes. The outmost circle indicates the hotspots associated with SBR response among seven models.

Gm16_29,491,946, Gm06_45,035,185, and Gm18_51,994,200, were identified to locate in close proximity to four main *P. pachyrhizi* R genes Rpp1, 2, 3, and 4, respectively, which were verified and reported in last decades.

Thirty-five candidate genes that might be associated with SBR disease resistance were found in the regions of the top 28 significant SNP markers (Table 1). Disease related annotations of these candidate genes were included but not limited to: LRR (Leucine Rich Repeat class protein), cytochrome 450, cell wall structure, RCC1 (regulator of chromosome condensation 1), AKR (ankyrin repeat-containing protein), F-box domain, NAC (NAM, ATAF and CUC family). Furthermore, most of the top 28 significant SNP regions were harboring more than one candidate gene, for example, the region of Gm02_7,235,181 and Gm02_7,234,594 contained three candidate genes, Glyma.02G083500, Glyma.02G083300, and Glyma.02G084100, coding for cell wall constituent, LRR-RLK, and RCC1, respectively.

Selection accuracy and selection efficiency

Selection accuracy (SA) and Selection efficiency (SE) reflect the contributions of selected alleles from the top 100 significant SNP to the resistance or susceptibility to *Phakopsora*. For the resistance alleles, SE varied from 50.0% to 84.2%, with an average of 57.5%; and the SA varied from 50.0% to 82.2%, with an average of 58.2%. SNP Gm09_1,951,644 had the highest values in both SA and SE in resistance effect. For susceptible alleles, the SE varied from 50.0% to 69.8%, with an average of 55.1%; and the SA varied from 50.3% to 56.9%, with an average of 52.7%. SNP Gm04_46,295,839 (52.7%) had the highest values in both SA and SE in susceptible effect (Table S6). This result identified the specific nucleotide of SBR-related alleles.

Genomic prediction

The 100 significant SNPs not only had the highest LOD value but were most repeatable across all GWAS methods as well. Following the same approach, six additional GWAS-based SNP sets were created, each consisting of 28, 100, 500, 1,000, 2,000, and 5,000 SNPs, respectively. In this study, we applied seven datasets, namely, All_SNPs (30,314), GWAS_5000SNPs, GWAS_2000SNPs, GWAS_1000SNPs, GWAS_500SNPs, GWAS_100SNPs and GWAS_28SNPs for GP analysis by five different GS models (Figure 5). The average GS accuracies of the All_SNPs set were at a medium level that was similar to those, ranging from 28.0% (RF) to 32.4% (gBLUP), among all the models.

Although the number of SNPs fluctuated by GWAS datasets, all the accuracy curves showed a similar pattern among the five models. The trend depicted by the left side of the curves indicated that as the number of SNPs decreases from 5,000 to 1,000, the accuracy of the prediction increases, too. The highest accuracies were observed when using the 1,000 SNP set, which were varying from 35.7% (RF) to 60.4% (Bayesian LASSO). And, as the number of SNPs continued to decrease from 1,000 to 100, the accuracy of GP also decreased. In all six GWAS based SNP sets, the Bayesian LASSO achieved the highest average GS accuracy of 53.0%, followed by rrBLUP with an average accuracy of 51.9%. On the other hand, the lowest accuracy of 36.2% was recorded when using the RF model. The GS accuracies of gBLUP and SVM models were at almost the same level but were relatively lower using the SVM model (Table S7).

10.3389/fpls.2023.1179357

TABLE 1 The genes within 50 kb genomic region of the top 28 significant SBR-associated SNPs with functional annotations.

SNP	GWAS model (Ranking)	LOD	Allele Type	Gene name	Functional annotations
Gm02_7235181	SUPER(1), FarmCPU, CMLM(5), MLMM(10)	7.91	T/C	Glyma.02G083500 Glyma.02G083300 Glyma.02G084100	LRR; RCC1; response to bacterial origin; defense response; structural constituent of cell wall
Gm02_7234594	SUPER(2), MLMM(11)	7.61	C/T		
Gm02_7315227	SUPER(3), GLM,MLM, Blink(5), MLMM(6)	7.52	G/A	Glyma.02G084100 Glyma.02G084900	RCC1 repeat; Ankyrin repeat family protein/ domain
Gm03_38913029	GLM, MLM, Blink (2), FarmCPU, CMLM(3), MLMM(7), GLM	6.85	T/C	Glyma.03G175800 Glyma.03G177400 Glyma.03G175300	Response to aluminum ion; cell wall; ABC transporter
Gm04_45884688	MLM, Blink(7), SUPER(15), MLMM (16), FarmCPU, CMLM(26)	6.23	T/C	Glyma.04g188000	LRR
Gm04_46003059	SUPER(20), MLMM(24)	6.03	G/A	Glyma.04G189300, Glyma.04g189500	Membrane; Cytochrome P450
Gm04_46295839	SUPER(16), MLMM(18)	6.08	C/T	Glyma.04G192300	Cell wall organization; cellular membrane fusion;
Gm04_46389651	SUPER(22), MLMM(27)	5.94	C/T		
Gm04_47132429	MLMM(4), FarmCPU, CMLM(6), GLM, MLM, Blink(13), SUPER(25)	5.78	T/C	Glyma.04G211100, Glyma.04G212000	NAC domain
Gm06_36808946	SUPER(6), GLM, MLM, Blink(9), FarmCPU, CMLM(34)	6.73	G/A	Glyma.06G232500	Response to molecule of bacterial origin
Gm08_43955878	FarmCPU, CMLM(19), SUPER(32), MLMM(33)	5.61	A/C	Glyma.08g319300, Glyma.08G321700	LRR; response to abscisic acid stimulus/cold/water deprivation
Gm09_1944730	MLMM(2), SUPER(27)	5.77	C/A	Glyma.09G024700	LRR-RLKs
Gm09_1943831	MLMM(3), SUPER(28)	5.73	G/A		
Gm09_1951644	FarmCPU, CMLM, MLMM (1), GLM, MLM, Blink (4),SUPER(18)	10.07	T/G		
Gm10_5573877	SUPER(5), MLMM(12), GLM, MLM, Blink(14)	6.73	C/T	Glyma.10G060100, Glyma.10G060200, Glyma.10G060600	Respiratory burst involved in defense response, response to bacterium/chitin; cell wall organization
Gm10_5573007	SUPER(7), MLMM(15)	6.58	C/T		
Gm10_5559592	SUPER(9), MLMM(20)	6.48	C/A		
Gm10_5541691	SUPER(33), MLMM(44)	5.60	C/T		
Gm10_5578693	SUPER(23), MLMM(32)	5.93	G/A		
Gm10_39142024	GLM, MLM, Blink(1), MLMM(8), SUPER(10), FarmCPU, CMLM(14)	7.12	C/T	Glyma.10g157500	LRR-RLKs, regulation of plant immunity
Gm10_39147121	MLMM(9), SUPER(21)	6.02	T/G		
Gm12_28136735	SUPER(4),GLM,MLM, Blink(8), MLMM(39)	7.03	G/A	Glyma.12G160100, Glyma.12G160400	NAC domain protein; Cytochrome P450
Gm13_16350701	FarmCPU, CMLM(16), GLM, MLM, Blink(23), SUPER(29)	5.63	T/C	Glyma.13G064500	F-box and WD40 domain protein, disease resistance protein
Gm14_2492139	GLM, MLM, Blink(6), SUPER(13), FarmCPU, CMLM(25), MLMM(26)	6.26	A/C	Glyma.14G034200, Glyma.14G040000	RCC1 family protein; LRR-RLKs
Gm14_6185611	MLMM(28), SUPER(36), GLM, MLM, Blink(46)	5.51	C/T	Glyma.14g073300, Glyma.14G073800	F-box domain; regulation of defense response
Gm16_4935328	GLM, MLM, Blink(10), MLMM(22), SUPER(31), FarmCPU, CMLM(32)	5.61	T/G	Glyma.16G051800, Glyma.16G052200	NAC domain protein; LRR-RLKs
Gm19_44734953	GLM, MLM, Blink(3), FarmCPU, CMLM(4), MLMM(25)	6.02	G/A	Glyma.19G189900, Glyma.19G190200, Glyma.19G190800	Defense response to bacterium; LRR-RLKs; plant-type cell wall
Gm20_36724867	FarmCPU, CMLM(2)	6.54	C/T	Glyma.20G124700	QSOX1 regulates plant immunity

Discussion

Phenotype

Resistance to *P. pachyrhizi* is commonly evaluated based on three types of SBR lesions: "TAN", "RB", and "Mixed". The "TAN" lesion type is characterized by heavy fungal sporulation typically develop on susceptible soybean leaves, while the RB or "reddish-brown" lesion type has been linked to resistance in known single gene resistance. The "Mixed" reaction is recorded when both RB and TAN lesions were observed on the same leaf (Miles et al., 2006). The simple classification of TAN and RB lesions had been widely used decades ago; however, it had been noted as oversimplified to the symptom observation. Nowadays, the appropriate practice is to separately divide TAN and RB into multiple classes to provide more accurate descriptions of disease symptoms while taking into account variations in fungal sporulation. Considering data consistency and method popularity, we took the TAN lesion as the phenotype of the association analysis for this study, which had sufficient observations and good distribution of SBR resistance. In the present study, the resistance resources were primarily sourced from China, Japan, and Korea, comprising 40%, 16%, and 21% of the total resources, respectively. These figures closely align with the respective proportions of 43%, 13%, and 18% observed in the overall population. In addition, according to the ANOVA between groups, it is obvious that the variability (99%) within groups is greater than the variability (1%) between groups (Table S8).

GWAS and candidate genes

Specific resistance to *P. pachyrhizi* is controlled by seven single dominant genes, namely, *Rpp1* (Chr 18), *Rpp2* (Chr16), *Rpp3* (Chr6), *Rpp4* (Chr7), *Rpp5* (Chr3), *Rpp6* (Chr18), and *Rpp7* (Chr19) (Calvo et al., 2008; Meyer et al., 2009; Lemos et al., 2011; Childs et al., 2018b). The single genes played an important role in SBR resistance, but this kind of resistance is not durable, and the usefulness of the sources loses its effectiveness once it is identified and applied in breeding (Chander et al., 2019). GWAS was performed in efforts to discover loci contributing SBR resistance, thus helping find all genes for SBR control (Chang et al., 2016). Multiple models were developed for GWAS based on linkage disequilibrium, including GLM, MLM, CMLM, MLMM, SUPER, FarmCPU, and BLINK (Wang and Zhang, 2021). Previous studies demonstrated that the differences of the models were caused by the interactions between the methods and other factors, including populations, sample size, mapping resolution, trait complexity, and quality of the data. Typically, all GWAS methods perform well when the aforementioned factors are favorable; however, each model may have varying numbers of false positives depends on the strengths and weaknesses of the model in different circumstance. Therefore, it is important to carefully consider the advantages and limitations of each GWAS method and choose the most appropriate one for the specific study and data. Additionally, multiple methods and independent replication studies are often used to confirm the validity of the results and minimize the risk of false positive findings. However, GWAS studies on SBR resistance were scarce, with the exception of a few studies that used a single model to discover loci contributing to general disease resistance in soybean (Kang et al., 2012; Chang et al., 2016). In this study, we applied all seven models and also considered both significance and consistency of each model for candidate SNPs of SBR resistance to hedge the false positives.

A total of four significant SNPs were located on or nearby the reported R genes. SNP Gm06_45,035,185 in chromosome 6 was located at gene *Rpp3*; Gm18_51,994,200 and Gm18_57,223,391 in chromosome 18 were nearby the genes *Rpp4*/*Rpp4-b* and *Rpp1*/*Rpp1-b*, respectively; and Gm16_29,491,946 in chromosome 16 was located at *Rpp2*, which showed the promise of GWAS on SBR resistance (Sharma and Gupta, 2006). However, we only observed moderate significance for these four SNPs in GWAS analysis, probably due to the following reasons: 1) different genetic variants contributing to the trait, rather than a single major gene; 2) major genes are often rare, the signal from a major gene may be diluted by underrepresented or even missing gene(s) in the samples.

Except for the major *Rpp*s, some significant SNPs also associate with *LRR* class genes that were considered to be the majority of disease resistance genes in plants (Kang et al., 2012). Genes encoding cytochrome P450 have been shown to contribute to both plant development and defense under pathogen attack (Siminszky et al., 1999). The F-box family proteins have been demonstrated to be directly involved in plant defense against pathogens(Liu and Xue, 2011). The QSOX1 (quiescin sulfhydryl oxidase homolog) were

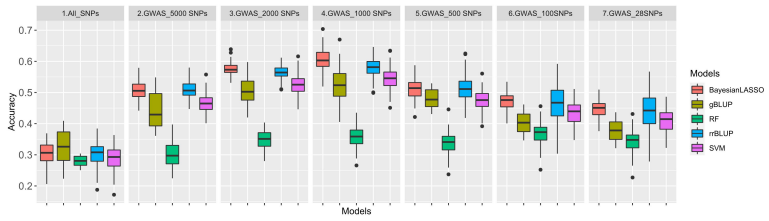


FIGURE 5
Genomic prediction (GP) accuracy for rust using five GP models: Ridge regression best linear unbiased predictor (rrBLUP) = blue, Genomic best linear unbiased predictor (gBLUP) = dark yellow, Bayesian least absolute shrinkage and selection operator(Bayesian LASSO) = red, Random Forest (RF) = green, Support vector machines (SVM) = purple based on seven datasets: All_SNPs (30314), and six GWAS based SNP sets with top28, 100, 500, 1,000, 2,000 and 5,000 SNPs.

reported to negatively regulate plant immunity against a pathogen (Chae et al., 2021); WD40 repeat-containing proteins which played an important effect on plant defense (Miller et al., 2016). The results were indicative of the robustness of the significant SNPs identified in this study. Other functional annotations pertaining to the candidate genes of cell wall structure/organization/construction and membrane fusion/proteins/structure/transporters have been demonstrated to play some roles in plant passive defense to pathogens (Mellersh and Heath, 2001; Hématy et al., 2009). The RCC1, NAC domain protein, ABC (ATP-binding cassette) transporters, etc. involve in many plant response-associated physiological activities to biotic or abiotic stresses and are widely annotated to the candidate genes (Table 1, S5) (Langenbach et al., 2016; Gautam et al., 2020; Oh et al., 2022). Furthermore, previous studies have reported the involvement of LRR (leucine-rich repeat), ABC transporters and F-box proteins in conferring resistance to rust fungi in other crop species belonging to the same order of *Pucciniales*, including wheat (Vikas et al., 2022), barley (Sallam et al., 2017), and maize (Juliana et al., 2018).

Selection accuracy and selection efficiency

SE and SA were computed for the significant SNPs associated SBR resistance or susceptibility (Ravelombola et al., 2017). The SA and SE of the marker were measured by relative proportion of an allele type (A/T/C/G) in resistant or/and susceptible accessions, as has been highlighted in other GWAS-related reports (Shi et al., 2016; Ravelombola et al., 2019). Specifically, the proportion of allele type for a completely un-associated SNP should be close to 50% in either resistant or susceptible group. Therefore, when the SA or SE value of the allele type is more than 50%, it contributes positively to the corresponding trait, or vice versa. In general, the two different nucleotides of any significant SNP must have opposite effects on disease resistance or susceptibility, which are defined as "R" or "S" alleles. We observed significant difference between "R" and "S" alleles in one SNP. For example, the "R" allele of SNP Gm04_46,295,839(C/T) has a "C" nucleotide with low SE and SA values (52.6% and 53.9%), but its "S" allele has a "T" nucleotide with high SE and SA values (67.8% and 57.%). This locus may relate to a S gene encoding a cellular membrane fusion protein as annotated in this study. On the contrary, the "A" allele of SNP Gm08_46,674,632(G/A) has high SE (84.2%) and SA (82.2%) values with resistance effect, whereas its "G" allele has low SE (51.5%) and SA (51.4%) values with susceptible effect. This locus is more likely to associate with a R gene coding for a LRR-containing protein in this study. In this study, all significant SNPs have higher than expected SA and SE values (>50%), suggesting that these SNPs can be used for further marker-assisted selection to enhance SBR resistance breeding in soybean.

Genomic prediction

The study discovered 28 significant SNPs located in 20 loci with genes that are associated with plant disease response or resistance. However, before applying these findings in breeding, further verification work is needed (Jannink et al., 2010; Crossa et al., 2017). GS has gained popularity in recent years in large-scale crop breeding programs. Previous studies have shown that GS achieves a more robust prediction of genotypic values compared to QTLs for traits controlled by many genes with small effects. GS tends to have a better and more reliable prediction than the traditional QTL approach, because it uses more markers that are distributed throughout the genome and captures more genetic variation of a trait (Bhat et al., 2016). GS can make predictions about an individual's performance even before it is phenotyped, which can save time and resources in the breeding process (Zhang et al., 2016; Ravelombola et al., 2019).

However, no research has investigated GS or GP for SBR resistance/tolerance. In this study, we performed GP with seven models on one All_SNP set and six GWAS-based SNP sets. The accuracies of All_SNP set (28.0%~32.4%) were similar to former studies on resistance/tolerance traits to abiotic and biotic stresses of several plant species, including wheat (Poland and Rutkoski, 2016), rice (Xu, 2013), maize (Technow et al., 2013), canola (Jan et al., 2016), alfalfa (Hawkins and Yu, 2018), cassava (Ly et al., 2013), oats (Asoro et al., 2011), miscanthus (Olatoye et al., 2020), grapevine (Brault et al., 2022), and intermediate wheatgrass (Crain et al., 2020). On the other hand, GWAS_SNPs-based GP accuracies were higher than those of All_SNP set-based, demonstrating the importance and contribution of significant SNPs in SBR resistance/tolerance. The accuracy using linear model gBLUP (45.5% in average) was close to those from machine learning (SVM), 47.5% in average, but lower than rrBLUP (51.2% in average) and Bayesian LASSO (52.0% in average) that had been considered to be the optimal approaches (Ravelombola et al., 2019).

Consistently with former reports (Bao et al., 2014; Li et al., 2018), we observed in this study that the accuracy of GP varied by the number of SNPs. For those GWAS-based SNP sets, a greater proportion of SNPs with more significance were retained for GS after further filtering of markers from 5,000 to 1,000, which led to increased accuracy. The accuracies of all models were improved until the number of SNPs reached 1,000, after which the accuracies began to decline until the number of markers dropped to 28. The apex of predictive accuracy was observed at a SNP count of 1,000, likely due to its ability to robustly capture LD and account for relatedness among soybean genotypes. An excess of SNPs beyond this threshold would introduce extraneous information to the models and elevate model complexity, while a SNP count lower than 1,000 would result in the loss of relevant information regarding LD and relatedness capture. Then again, the objective of this GWAS study was primarily to identify the associated loci and candidate genes related to SBR. The use of multiple SNP sets and GS models was employed to ensure the consistency of the GWAS results, rather than to quantitatively evaluate the superiority or variations between the models and data sets. However, the above results can still serve as a reference for future GS research in disease resistance.

Phenotypic selection has been successfully implemented for disease resistance, but without controlled experiments, it is difficult to determine whether the resistance is quantitative or qualitative. Therefore, it is difficult to determine whether the resistance will be durable in the long term. In this study, the SBR-related markers we identified can be used to select for both quantitative and qualitative disease resistance within the breeding lines to bypass the need for

controlled experiments through the use of conventional MAS. Additionally, by utilizing GP, the breeders can select for the accumulation of QTL associated with resistance, thereby taking advantage of both quantitative and qualitative resistance genes, even those that have not yet been characterized. This allows them to select the most promising lines for further development and testing without multiple generations of phenotyping.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Author contributions

HX, AS, JW, and YC organized and analysed the original data. HX and Y-BP drafted the manuscript. WL, Y-BP and AS critically revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by USDA NIFA HATCH project ARK0VG2018 and a USDA-ARS Non-Assistance Cooperative Agreement on Genetic Analysis and Trait-Specific Molecular Marker Development (Accession No. 440501).

Acknowledgments

The authors are thankful to Reid D. Frederick, Glen L. Hartman, and Monte R. Miles, the USDA-ARS, Foreign Disease-Weed Science Research Unit, Urbana, IL, USA for publicly accessible SBR disease scores and phenotyping data from the USDA GRIN website. The authors also thank Zhongqi He and Yulin Jia for review comments. USDA is an equal opportunity provider and employer.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1179357/full#supplementary-material

SUPPLEMENTARY FIGURE 1
The linkage disequilibrium decay rate was estimated as squared correlation coefficient (r^2) using all pairs of SNPs located within 4 Mb of physical distance in euchromatic. The red line is the moving average of the (r2) value of the ten adjacent markers.

SUPPLEMENTARY FIGURE 2
The Manhattan plots for SBR by multi-GWAS models: **(A)** Blink, **(B)** GLM, **(C)** MLM, **(D)** CMLM, **(E)** FarmCPU, **(F)** SUPER, **(G)** MLMM. Additionally: **(H)** QQ-plots of the above seven models.

References

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., and Jannink, J. L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite north American oats. *Plant Genome* 4, 132–144. doi: 10.3835/plantgenome2011.02.0007

Bao, Y., Vuong, T., Meinhardt, C., Tiffin, P., Denny, R., Chen, S., et al. (2014). Potential of association mapping and genomic selection to explore PI 88788 derived soybean cyst nematode resistance. *Plant Genome* 7, 2840–2854. doi: 10.3835/plantgenome2013.11.0039

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., et al. (2016). Genomic Selection in the Era of Next Generation Sequencing for Complex Traits in Plant Breeding. *Front Genet.* 7, 221–2854. doi: 10.3389/fgene.2016.00221

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brault, C., Segura, V., This, P., Le Cunff, L., Flutre, T., François, P., et al. (2022). Across-population genomic prediction in grapevine opens up promising prospects for breeding. *Horticult. Res.* 9, uhac041. doi: 10.1093/hr/uhac041

Bromfield, K., and Hartwig, E. (1980). Resistance to soybean rust and mode of inheritance. *Crop Sci.* 20, 254–255. doi: 10.2135/cropsci1980.0011183X002000020026x

Büschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., et al. (1997). The barley mlo gene: a novel control element of plant pathogen resistance. *Cell* 88, 695–705. doi: 10.1016/S0092-8674(00)81912-1

Calvo, É. S., Kiihl, R. A., Garcia, A., Harada, A., and Hiromoto, D. M. (2008). Two major recessive soybean genes conferring soybean rust resistance. *Crop Sci.* 48, 1350–1354. doi: 10.2135/cropsci2007.10.0589

Chae, H. B., Kim, M. G., Kang, C. H., Park, J. H., Lee, E. S., Lee, S.-U., et al. (2021). Redox sensor QSOX1 regulates plant immunity by targeting GSNOR to modulate ROS generation. *Mol. Plant* 14, 1312–1327. doi: 10.1016/j.molp.2021.05.004

Chander, S., Ortega-Beltran, A., Bandyopadhyay, R., Sheoran, P., Ige, G. O., Vasconcelos, M. W., et al. (2019). Prospects for durable resistance against an old soybean enemy: a four-decade journey from Rpp1 (Resistance to phakopsora pachyrhizi) to Rpp7. *Agronomy* 9, 348. doi: 10.3390/agronomy9070348

Chang, H.-X., Lipka, A. E., Domier, L. L., and Hartman, G. L. (2016). Characterization of disease resistance loci in the USDA soybean germplasm collection using genome-wide association studies. *Phytopathology* 106, 1139–1151. doi: 10.1094/PHYTO-01-16-0042-FI

Childs, S. P., Buck, J. W., and Li, Z. (2018a). Breeding soybeans with resistance to soybean rust (Phakopsora pachyrhizi). *Plant Breed.* 137, 250–261. doi: 10.1111/pbr.12595

Childs, S. P., King, Z. R., Walker, D. R., Harris, D. K., Pedley, K. F., Buck, J. W., et al. (2018b). Discovery of a seventh rpp soybean rust resistance locus in soybean accession PI 605823. *Theor. Appl. Genet.* 131, 27–41. doi: 10.1007/s00122-017-2983-4

Crain, J., Bajgain, P., Anderson, J., Zhang, X., Dehaan, L., and Poland, J. (2020). Enhancing crop domestication through genomic selection, a case study of intermediate wheatgrass. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00319

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

De Wit, P. J. (1992). Molecular characterization of gene-for-gene systems in plant-fungus interactions and the application of avirulence genes in control of plant pathogens. *Annu. Rev. Phytopathol.* 30, 391–418. doi: 10.1146/annurev.py.30.090192.002135

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250-255. doi: 10.3835/plantgenome2011.08.0024

Frichot, E., and François, O. (2015). LEA: an R package for landscape and ecological association studies. *Methods Ecol. Evol.* 6, 925–929. doi: 10.1111/2041-210X.12382

Garcia, A., Calvo, É. S., De Souza Kiihl, R. A., Harada, A., Hiromoto, D. M., and Vieira, L. G. E. (2008). Molecular mapping of soybean rust (Phakopsora pachyrhizi) resistance genes: discovery of a novel locus and alleles. *Theor. Appl. Genet.* 117, 545–553. doi: 10.1007/s00122-008-0798-z

Gautam, A., Pandey, A. K., and Dubey, R. S. (2020). Unravelling molecular mechanisms for enhancing arsenic tolerance in plants: a review. *Plant Gene* 23, 100240. doi: 10.1016/j.plgene.2020.100240

Gebremedhn, H. M., Msiska, U. M., Weldekidan, M. B., Odong, T. L., Rubaihayo, P., and Tukamuhabwa, P. (2020). Prediction of candidate genes associated with resistance to soybean rust (Phakopsora pachyrhizi) in line UG-5. *Plant Breed.* 139, 943–949. doi: 10.1111/pbr.12847

Godoy, C., Meyer, M.Braunschweig: Deutsche Phytomedizinische Gesellschaft (2020). "Overcoming the threat of Asian soybean rust in Brazil," in *Modern fungicides antifungal compounds, IX, ed.* Eds. B. Fraaije, H. B. Deising, A. Mehl, E. C. Oerke, H. Sierotzki and G. Stammler (Braunschweig: Deutsche Phytomedizinische Gesellschaft), 51–56.

Godoy, C. V., Seixas, C. D. S., Soares, R. M., Marcelino-Guimarães, F. C., Meyer, M. C., and Costamilan, L. M. (2016). Asian Soybean rust in Brazil: past, present, and future. *Pesquisa Agropecuária Bras.* 51, 407–421. doi: 10.1590/S0100-204X2016000500002

Hartman, G. L., Miles, M. R., and Frederick, R. D. (2005). Breeding for resistance to soybean rust. *Plant Dis.* 89, 664–666. doi: 10.1094/PD-89-0664

Hartman, G., Wang, T., and Tschanz, A. (1991). Soybean rust development and the quantitative relationship between rust severity and soybean yield. *Plant Disease* 75 (6), 596-600. doi: 10.1094/PD-75-0596

Hawkins, C., and Yu, L.-X. (2018). Recent progress in alfalfa (Medicago sativa l.) genomics and genomic selection. *Crop J.* 6, 565–575. doi: 10.1016/j.cj.2018.01.006

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., and Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00484

Hématy, K., Cherk, C., and Somerville, S. (2009). Host–pathogen warfare at the plant cell wall. *Curr. Opin. Plant Biol.* 12, 406–413. doi: 10.1016/j.pbi.2009.06.007

Heslot, N., Yang, H. P., Sorrells, M. E., and Jannink, J. L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Isard, S. A., Gage, S. H., Comtois, P., and Russo, J. M. (2005). Principles of the atmospheric pathway for invasive species applied to soybean rust. *Bioscience* 55, 851–861. doi: 10.1641/00063568(2005)055[0851:POTAPF]2.0.CO;2

Jan, H. U., Abbadi, A., Lücke, S., Nichols, R. A., and Snowdon, R. J. (2016). Genomic prediction of testcross performance in canola (Brassica napus). *PloS One* 11, e0147769. doi: 10.1371/journal.pone.0147769

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings Funct. Genomics* 9, 166–177. doi: 10.1093/bfgp/elq001

Jiang, J., and Nguyen, T. (2021). *Linear and generalized linear mixed models and their applications* (New York, NY: Springer Nature).

Juliana, P., Singh, R. P., and Singh, P. K. (2018). Genome-wide association mapping for resistance to leaf rust, stripe rust and tan spot in wheat reveals potential candidate genes. *Theor. Appl. Genet.* 131, 1405–1422. doi: 10.1007/s00122-018-3086-6

Kang, Y. J., Kim, K. H., Shim, S., Yoon, M. Y., Sun, S., Kim, M. Y., et al. (2012). Genome-wide mapping of NBS-LRR genes and their association with disease resistance in soybean. *BMC Plant Biol.* 12, 1–13. doi: 10.1186/1471-2229-12-139

Kashiwa, T., Muraki, Y., and Yamanaka, N. (2020). Near-isogenic soybean lines carrying Asian soybean rust resistance genes for practical pathogenicity validation. *Sci. Rep.* 10, 1–7. doi: 10.1038/s41598-020-70188-7

Langenbach, C., Campe, R., Beyer, S. F., Mueller, A. N., and Conrath, U. (2016). Fighting Asian soybean rust. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00797

Lemos, N. G., Braccini, A. D. L. E., Abdelnoor, R. V., De Oliveira, M. C. N., Suenaga, K., and Yamanaka, N. (2011). Characterization of genes Rpp2, Rpp4, and Rpp5 for resistance to soybean rust. *Euphytica* 182, 53–64. doi: 10.1007/s10681-011-0465-3

Levy, C. (2005). Epidemiology and chemical control of soybean rust in southern Africa. *Plant Dis.* 89, 669–674. doi: 10.1094/PD-89-0669

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., et al. (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00190

Li, S., Smith, J. R., Ray, J. D., and Frederick, R. D. (2012). Identification of a new soybean rust resistance gene in PI 567102B. *Theor. Appl. Genet.* 125, 133–142. doi: 10.1007/s00122-012-1821-y

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PloS Genet.* 12, e1005767. doi: 10.1371/journal.pgen.1005767

Liu, T.-B., and Xue, C. (2011). The ubiquitin-proteasome system and f-box proteins in pathogenic fungi. *Mycobiology* 39, 243–248. doi: 10.5941/MYCO.2011.39.4.243

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., et al. (2011). Genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110, 77–123. doi: 10.1016/B978-0-12-385531-2.00002-5

Lucas, J. A. (2020). *Plant pathology and plant pathogens* (UK: Wiley-Blackwell).

Ly, D., Hamblin, M., Rabbi, I., Melaku, G., Bakare, M., Gauch, H. G.Jr., et al. (2013). Relatedness and genotype× environment interaction affect prediction accuracies in genomic selection: a study in cassava. *Crop Sci.* 53, 1312–1325. doi: 10.2135/cropsci2012.11.0653

Mellersh, D. G., and Heath, M. C. (2001). Plasma membrane–cell wall adhesion is required for expression of plant defense responses during fungal penetration. *Plant Cell* 13, 413–424. doi: 10.1105/tpc.13.2.413

Meyer, J. D., Silva, D. C., Yang, C., Pedley, K. F., Zhang, C., Van De Mortel, M., et al. (2009). Identification and analyses of candidate genes for Rpp4-mediated resistance to Asian soybean rust in soybean. *Plant Physiol.* 150, 295–307. doi: 10.1104/pp.108.134551

Miles, M. R., Frederick, R. D., and Hartman, G. L. (2003). "Soybean rust: is the US soybean crop at risk," in *APS Net Plant Pathology Online*. Available at: https://www.ars.usda.gov/research/publications/publication/?seqNo115=150029.

Miles, M., Frederick, R., and Hartman, G. (2006). Evaluation of soybean germplasm for resistance to phakopsora pachyrhizi. *Plant Health Prog.* 7, 33. doi: 10.1094/PHP-2006-0104-01-RS

Miller, J. C., Chezem, W. R., and Clay, N. K. (2016). Ternary WD40 repeat-containing protein complexes: evolution, composition and roles in plant immunity. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01108

Ogutu, J. O., Piepho, H. P., and Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* 5 (Suppl 3), S11. doi: 10.1186/1753-6561-5-S3-S11

Oh, Y., Lee, S., Rioux, R., Singh, P., Jia, M. H., Jia, Y., et al. (2022). Analysis of differentially expressed rice genes reveals the ATP-binding cassette (ABC) transporters as a candidate gene against the sheath blight pathogen, rhizoctonia solani. *Phytofrontiers* 2, 105–115. doi: 10.1094/PHYTOFR-05-21-0035-R

Olatoye, M. O., Clark, L. V., Labonte, N. R., Dong, H., Dwiyanti, M. S., Anzoua, K. G., et al. (2020). Training population optimization for genomic selection in miscanthus. *G3: Genes Genomes Genet.* 10, 2465–2476. doi: 10.1534/g3.120.401402

Pandey, A. K., Yang, C., Zhang, C., Graham, M. A., Horstman, H. D., Lee, Y., et al. (2011). Functional analysis of the Asian soybean rust resistance pathway mediated by Rpp2. *Mol. Plant-Microbe Interact.* 24, 194–206. doi: 10.1094/MPMI-08-10-0187

Pivonia, S., and Yang, X. (2004). Assessment of the potential year-round establishment of soybean rust throughout the world. *Plant Dis.* 88, 523–529. doi: 10.1094/PDIS.2004.88.5.523

Poland, J. A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5, 103–113. doi: 10.3835/plantgenome2012.06.0006

Poland, J., and Rutkoski, J. (2016). Advances and challenges in genomic selection for disease resistance. *Annu. Rev. Phytopathol.* 54, 79–98. doi: 10.1146/annurev-phyto-080615-100056

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559-575. doi: 10.1086/519795

Ravelombola, W., Qin, J., Shi, A., Lu, W., Weng, Y., Xiong, H., et al. (2017). Association mapping revealed SNP markers for adaptation to low phosphorus conditions and rock phosphate response in USDA cowpea (Vigna unguiculata (L.) walp.) germplasm. *Euphytica* 213, 1–14. doi: 10.1007/s10681-017-1971-8

Ravelombola, W. S., Qin, J., Shi, A., Nice, L., Bao, Y., Lorenz, A., et al. (2019). Genome-wide association study and genomic selection for soybean chlorophyll content associated with soybean cyst nematode tolerance. *BMC Genomics* 20, 1–18. doi: 10.1186/s12864-019-6275-z

Revell, L. J. (2012). Phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. doi: 10.1111/j.2041-210X.2011.00169.x

Sallam, A. H., Tyagi, P., Brown-Guedira, G., Muehlbauer, G. J., Hulse, A., and Steffenson, B. J. (2017). Genome-wide association mapping of stem rust resistance in

hordeum vulgare subsp. spontaneum. *G3 Genes Genomes Genet.* 7 (10), 3491–3507. doi: 10.1534/g3.117.300222

Sharma, S., and Gupta, G. (2006). Current status of soybean rust (Phakopsora pachyrhizi)-a review. *Agric. Rev.* 27, 91–102. Available at: https://www.indianjournals.com/ijor.aspx?target=ijor:ar&volume=27&issue=2&article=002.

Shi, A., Buckley, B., Mou, B., Motes, D., Morris, J. B., Ma, J., et al. (2016). Association analysis of cowpea bacterial blight resistance in USDA cowpea germplasm. *Euphytica* 208, 143–155. doi: 10.1007/s10681-015-1610-1

Shikha, M., Kanika, A., Rao, A. R., Mallikarjuna, M. G., Gupta, H. S., and Nepolean, T. (2017). Genomic selection for drought tolerance using genome-wide SNPs in maize. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00550

Siminszky, B., Corbin, F. T., Ward, E. R., Fleischmann, T. J., and Dewey, R. E. (1999). Expression of a soybean cytochrome P450 monooxygenase cDNA in yeast and tobacco enhances the metabolism of phenylurea herbicides. *Proc. Natl. Acad. Sci.* 96, 1750–1755. doi: 10.1073/pnas.96.4.1750

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PloS One* 8 (1), e54985. doi: 10.1371/journal.pone.0054985

Technow, F., Bürger, A., and Melchinger, A. E. (2013). Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes Genomes Genet.* 3, 197–203. doi: 10.1534/g3.112.004630

Uppalapati, S. R., Ishiga, Y., Doraiswamy, V., Bedair, M., Mittal, S., Chen, J., et al. (2012). Loss of abaxial leaf epicuticular wax in medicago truncatula irg1/palm1 mutants results in reduced spore differentiation of anthracnose and nonhost rust pathogens. *Plant Cell* 24, 353–370. doi: 10.1105/tpc.111.093104

Vikas, V. K., Pradhan, A. K., and Budhlakoti, N. (2022). Multi-locus genome-wide association studies (ML-GWAS) reveal novel genomic regions associated with seedling and adult plant stage leaf rust resistance in bread wheat (Triticum aestivum l.). *Heredity* 128, 434–449. doi: 10.1038/s41437-022-00525-1

Visscher, P. M., Brown, M. A., Mccarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Waheed, S., Anwar, M., Saleem, M. A., Wu, J., Tayyab, M., and Hu, Z. (2021). The critical role of small RNAs in regulating plant innate immunity. *Biomolecules* 11, 184. doi: 10.3390/biom11020184

Walker, D., Boerma, H., Phillips, D., Schneider, R., Buckley, J., Shipe, E., et al. (2011). Evaluation of USDA soybean germplasm accessions for resistance to soybean rust in the southern united states. *Crop Sci.* 51, 678–693. doi: 10.2135/cropsci2010.06.0340

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., and Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PloS One* 9, e107684. doi: 10.1371/journal.pone.0107684

Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinf* 19, 629-640. doi: 10.1016/j.gpb.2021.08.005

Wen, Y.-J., Zhang, H., Ni, Y.-L., Huang, B., Zhang, J., Feng, J.-Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings Bioinf.* 19, 700–712. doi: 10.1093/bib/bbw145

Xu, S. (2013). Genetic mapping and genomic selection using recombination breakpoint data. *Genetics* 195, 1103–1115. doi: 10.1534/genetics.113.155309

Yorinori, J., Paiva, W., Frederick, R., Costamilan, L., Bertagnolli, P., Hartman, G., et al. (2005). Epidemics of soybean rust (Phakopsora pachyrhizi) in Brazil and Paraguay from 2001 to 2003. *Plant Dis.* 89, 675–677. doi: 10.1094/PD-89-0675

Zhang, J., Song, Q., Cregan, P. B., and Jiang, G.-L. (2016). Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (Glycinemax). *Theor. Appl. Genet.* 129, 117–130. doi: 10.1007/s00122-015-2614-x

Zhang, Z., Todhunter, R., Buckler, E., and Van Vleck, L. D. (2007). Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *J. Anim. Sci.* 85, 881–885. doi: 10.2527/jas.2006-656

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 30 May 2023
DOI 10.3389/fpls.2023.1135884

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Pritam Kalia,
Indian Agricultural Research Institute (ICAR), India
Rafael Fernández-Muñoz,
Spanish National Research Council (CSIC), Spain
Md Abdur Rahim,
Sher-e-Bangla Agricultural University, Bangladesh

*CORRESPONDENCE
Dilip R. Panthee
dilip_panthee@ncsu.edu

RECEIVED 02 January 2023
ACCEPTED 02 May 2023
PUBLISHED 30 May 2023

CITATION
Adhikari TB, Siddique MI, Louws FJ, Sim S-C and Panthee DR (2023) Molecular mapping of quantitative trait loci for resistance to early blight in tomatoes.
Front. Plant Sci. 14:1135884.
doi: 10.3389/fpls.2023.1135884

Molecular mapping of quantitative trait loci for resistance to early blight in tomatoes

Tika B. Adhikari[1], Muhammad Irfan Siddique[2], Frank J. Louws[1,3], Sung-Chur Sim[4] and Dilip R. Panthee[2]*

[1]Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, United States, [2]Department of Horticultural Science, North Carolina State University, Mountain Horticultural Crops Research and Extension Center, Mills River, NC, United States, [3]Department of Horticultural Science, North Carolina State University, Raleigh, NC, United States, [4]Department of Bioresources Engineering, Sejong University, Seoul, Republic of Korea

Early blight (EB), caused by *Alternaria linariae* (Neerg.) (syn. *A. tomatophila*) Simmons, is a disease that affects tomatoes (*Solanum lycopersicum* L.) throughout the world, with tremendous economic implications. The objective of the present study was to map the quantitative trait loci (QTL) associated with EB resistance in tomatoes. The F_2 and $F_{2:3}$ mapping populations consisting of 174 lines derived from NC 1CELBR (resistant) × Fla. 7775 (susceptible) were evaluated under natural conditions in the field in 2011 and in the greenhouse in 2015 by artificial inoculation. In all, 375 Kompetitive Allele Specific PCR (KASP) assays were used for genotyping parents and the F_2 population. The broad-sense heritability estimate for phenotypic data was 28.3%, and 25.3% for 2011, and 2015 disease evaluations, respectively. QTL analysis revealed six QTLs associated with EB resistance on chromosomes 2, 8, and 11 (LOD 4.0 to 9.1), explaining phenotypic variation ranging from 3.8 to 21.0%. These results demonstrate that genetic control of EB resistance in NC 1CELBR is polygenic. This study may facilitate further fine mapping of the EB-resistant QTL and marker-assisted selection (MAS) to transfer EB resistance genes into elite tomato varieties, including broadening the genetic diversity of EB resistance in tomatoes.

KEYWORDS

early blight, heritability estimates, QTL analysis, tomatoes, *Solanum lycopersicum* (L.)

Introduction

Early blight (EB), caused by *Alternaria linariae* (Neerg.) (syn. *A. tomatophila*) Simmons, once classified within *A. solani*), is a serious threat to tomato-producing areas across the globe and particularly in the Southeast USA (Nash and Gardner, 1988). EB symptoms are typically characterized by the formation of dark necrotic lesions with concentric rings on the leaves. Consequently, blighted leaves are defoliated, which can

reduce fruit quality and yield (Basu, 1974; Jones, 1991; Rotem, 1994). Due to a lack of cultivars with efficacious resistance, tomato growers have relied on other control measures, such as field sanitation, crop rotation, cultural practices, and intensive calendar-based fungicide application programs (Gleason et al., 1995; Keinath et al., 1996; Louws et al., 1996). One of the strategies to manage EB in tomatoes is the frequent application of quinone-oxidizing inhibitors (Q_oI; strobilurins), such as azoxystrobin and pyraclostrobin (a single site mode of action fungicide), or protectant fungicides, such as mancozeb and chlorothalonil (Ivors et al., 2007). In potato fields, a shift of *A. linariae* isolates toward Q_oI fungicide resistance has been reported due to the F129L mutation (Pasche et al., 2005; Pasche and Gudmestad, 2008), and resistant strains have been confirmed in NC (Inga Meadow, personal communication). In the past decades, three EB forecast systems have been developed and used to curtail the costs of and to optimize disease management (Madden et al., 1978; Pennypacker et al., 1983; Pitblado, 1992; Gleason et al., 1995; Keinath et al., 1996; Louws et al., 1996; Cowgill et al., 2005). Among the disease forecasting systems, Tomato Disease Forecaster (TOM-CAST) was deemed an effective strategy to determine the proper timing of fungicide sprays (Pitblado, 1992).

While the use of fungicides can manage EB, it is preferred to grow a resistant variety to manage the disease. So far, no single-gene conferring resistance to EB has been identified in the cultivated tomato or its wild relatives (Zhang et al., 2003). Although a great deal of effort has been made toward developing tomato cultivars resistant to EB at North Carolina State University (NCSU), only a few moderately resistant lines and cultivars have been identified (Gardner, 1984; Gardner, 1988; Nash and Gardner, 1988; Adhikari et al., 2017). These tomato lines and cultivars exhibited partial resistance to EB under severe epidemics but were either late maturing or low-yielding (Foolad et al., 2002; Zhang et al., 2003). In many cases, resistance to EB in tomatoes has been reported to be a complex trait and controlled by quantitative and partially dominant genes with epitasis (Gardner, 1988; Nash and Gardner, 1988; Gardner and Shoemaker, 1999; Gardner and Panthee, 2012). To resolve these problems, quantitative trait loci (QTL) mapping can serve as a suitable approach to unraveling the genetic control of complex and polygenic traits in segregating populations and can provide valuable information on phenotypic trait–molecular marker associations (Wurschum, 2012).

In the past, different molecular markers have been used to identify QTL for EB resistance and to develop consensus genetic maps in tomatoes. Among these, restriction fragment length polymorphisms (RFLPs), microsatellites or simple sequence repeats (SSRs), and resistance gene analogs (RGAs) have been widely used to identify specific genomic regions associated with resistance to EB (Foolad et al., 2002; Zhang et al., 2003; Chaerani et al., 2007; Adhikari et al., 2017). The development of single nucleotide polymorphisms (SNP) molecular markers (Jiménez-Gómez and Maloof, 2009), which are the most abundant source of variation in the genome for both intragenic and intergenic regions, represents a valuable tool to identify polymorphisms among closely related lines and to develop highly saturated genetic maps (Sim et al., 2012b).

In this study, the 174 F_2-derived F_3 ($F_{2:3}$) population, from a cross between the resistant tomato line NC 1CELBR and the susceptible tomato cultivar Fla. 7775, was phenotyped for EB resistance in the field and under controlled conditions in the greenhouse and genotyped with single nucleotide polymorphism (SNP) molecular markers. QTL analysis was performed to identify the putative genomic regions associated with resistance to EB in the tomato.

Materials and methods

Plant materials

Tomato breeding line NC 1CELBR was developed at North Carolina State University (NCSU). It is a large-fruited fresh-market tomato line with determinate growth habits and is resistant to EB. The line was developed by multiple crosses involving wild species *S. habrochaites* and *S. pimpinellifolium* (Gardner and Panthee, 2010). Dr. Jay Scott, University of Florida, kindly provided the seed of the susceptible cultivar Fla. 7775. Despite other similar characteristics, contrasting EB reactions in NC 1CELBR and Fla. 7775 provided ideal materials to develop a population for genetic mapping studies. Crosses were made in the fall of 2009 at the Mountain Horticultural Crops Research and Extension Center (MHCREC), (NCSU), Mills River, North Carolina (NC). The F_2 seeds were produced in the spring of 2010 by selfing the F_1. Subsequently, 174 $F_{2:3}$ families were developed and used for phenotypic evaluation in the field and greenhouse, SNP marker analysis, and QTL mapping.

Phenotyping of the F_2 population in the field in 2011 in Waynesville, NC

To evaluate plants for resistance to EB in the field, the experiment was conducted in 2011. Seeds were planted in 72 cell flats (56 × 28 cm^2) in potting mix in the first week of May, and transplants at about six weeks from seed were planted. In the first week of June 2011, greenhouse-grown seedlings of the 174 F_2 and F_1 hybrid (NC 10175), susceptible controls (Fletcher, NC123S and NC 30P), resistant controls (NC 2CELBR and Mountain Merit), and resistant and susceptible parents (NC 1CELBR and Fla. 7775) were planted at the Mountain Research Station, Waynesville, NC. Spacing was 45 cm between plant-to-plant and 150 cm between row-to-row. The soil was a clay-loam texture, and the natural daylight photoperiod was about 14/10 hr, with temperatures averaging 25-30°C at their high and 14-16°C at their low. This field site was chosen because *A. linariae* inoculum naturally occurs each year almost three weeks after transplanting. Parents and F_1 were planted as a control to make sure that the disease developed well in the susceptible parent and that the resistant parent was healthy even under high inoculum pressure. No fungicide application was made to control the EB whereas late blight and Septoria leaf spot-specific fungicides were applied to control those diseases by spraying Presidio every week in combination with others as per the fungicide spray guide in NC (Ivors, 2011). Each

plant was assessed for EB symptoms six weeks after planting to the field using a Horsfall and Barratt (1945) rating scale of 1 to 11, where 1 indicates no EB symptoms on the leaf surface, and 11 indicates complete defoliation. Humid and warm conditions favor *A. linariae* development, which was conducive to EB development in 2011.

Phenotyping of the $F_{2:3}$ population in the greenhouse in 2015 in Mills River, NC

Seeds of the 174 $F_{2:3}$ population and resistant and susceptible parents (NC 1CELBR and Fla. 7775) were surface-sterilized and sown in the greenhouse at Mills River. Seeds were sown in 4P soil mixture (Fafard®, Florida, USA) in flatbed metal trays in a standard seeding mix (2:2:1 v/v/v) peat moss: pine bark: vermiculite with macro- and micro-nutrients (Van Wingerden International Inc., Mills River, NC) in March 2015. After ten days, seedlings were transplanted to 24-cell flats (56 cm x 28cm). Three plants per genotype were planted with two replications, and the experiment was conducted in a completely randomized design. Plants in the greenhouse study were fertilized using a 20:20:20 ratio of nitrogen, phosphorus, and potassium, respectively. Standard greenhouse pesticide application was used for possible insect and bacterial disease control. A single-spore isolate of *A. linariae* Sorauer collected from naturally infected tomato plants in Hendersonville, NC was used in this study. The fungus was isolated from infected leaf tissues and grown on potato dextrose agar (PDA, 39 g of Difco PDA, Becton, Dickinson and Company, Sparks, MD) in 10-cm Petri dishes and incubated at 23° C under white fluorescent lamps with a 12-h photoperiod. This isolate collected from the field was confirmed as *A. solani* using microscopic examination and PCR-based assays (Gannibal et al., 2014). After 10-12 days, conidia were harvested by flooding the plates with sterile distilled water. The inoculum concentration was adjusted to 1×10^7 conidia mL^{-1} using a hemocytometer. Before inoculation, a drop (~10 µL) of Tween 20 (Polyoxyethylene-20-sorbitan monolaurate) was added to the inoculum suspension to facilitate uniform spore deposition onto leaves. Nine-week-old plants were artificially inoculated using a hand sprayer (R & D Sprayers Inc., Opelousas, LA, USA). After inoculation, plants were placed in the dark for 24 h and covered entirely with white plastic to create a relative humidity of > 95%. Each inoculated plant was scored for EB symptoms using a Horsfall-Barratt rating scheme (Horsfall and Barratt, 1945) at 14 and 21 days after inoculation, as described above. Average disease scores were used to measure resistance to EB and to identify QTL in the greenhouse trials.

DNA isolation and SNP genotyping

Genomic DNA of young leaf tissues of each parent and individual plant from F_2 generation was extracted using the DNeasy Plant Mini Kit (Qiagen Inc., Valencia, CA). A NanoDrop (Model ND-2000, Thermo Scientific Inc., Wilmington, DE) was used to quantify each DNA sample. Approximately, 50 ng/µl of DNA was prepared from each sample for SNP genotyping. We used an optimized subset of 384 SNPs markers that were derived from the 7,725 SNP array developed by the Solanaceae Coordinated Agricultural Project (SolCAP) (Sim et al., 2012a; Sim et al., 2012b). The subset of markers was selected based on polymorphism rates among six fresh market tomato accessions, including Fla.7776, Fla. 8383, NC33EB-1, 091120-7, Fla. 7775, and NC 1CELBR. Also, the genetic position in the genome based on recombination (Sim et al., 2012a) and the physical position was considered important selection criteria to ensure genome coverage. These 384 SNPs were analyzed using the Kompetitive Allele Specific PCR (KASP) genotyping platform (LGC Genomics, Beverly, MA).

Data analysis

The visual illustration of the correlation matrix and principal component analysis (PCA) was done by using the R language v3.2.3 coupled with the RStudio interface v1.0.143 and R packages ("FactoMineR", "factoextra", "ggplot2", "ggplots", "corrplot"), respectively (R Core Team, 2018; Amanullah et al., 2022). The summary statistics and normal probability plots were calculated using the UNIVARIATE procedure of SAS. The heritability was estimated for each environment by calculating variance components using the 'ASYCOV' function in PROC MIXED in SAS (SAS Institute Inc., 2012).

Broad-sense heritability (H^2) was estimated using the following variance components from the F_2 population (Nyquist, 1991; Falconer and Mackay, 1996):

$$H^2 = \frac{VG}{VP} = \frac{VA + VD}{VA + VD + VE}$$

Narrow-sense heritability (h^2) was determined using a regression analysis of offspring on parent approach, using data from the F_2 and F_3 generations as has been used by Ohlson and Foolad (2015) and as follows (Nyquist, 1991; Falconer and Mackay, 1996):

$$h^2 = \frac{VA}{VA + VD + VE} = \frac{Cov(F3xF2)}{\sqrt{(VF3xVF2)}}$$

Where, H= broad-sense heritability, h^2= narrow-sense heritability, VG=genetic variance, VP=phenotypic variance, VA = additive variance, VD= dominance variance, VE=error variance, VF_2 = Variance at F_2 generation, VF_3 = Variance at F_3 generation, and Cov (F_3xF_2) = Covariance of individuals at F_2 and F_3 generations.

Linkage map construction of F_2 and QTL analysis

Of the 384 SNP markers tested, 375 were polymorphic between the two parental lines, NC 1CELBR and Fla. 7775, that were used for genetic map construction (Meng et al., 2015). The linkage map was constructed using JoinMap 4.0 (van Ooijen, 2006). The grouping mode was set as the autonomous limit of detection, the mapping algorithm was used to perform regression mapping (limit of detection > 2.5, recombination frequency< 0.4, and jump = 5)

(Asekova et al., 2021). The Kosambi mapping function was used to convert recombination frequencies into map distance (Kosambi, 1943). Independent limit of detection and maximum likelihood algorithms were used for grouping and ordering of markers, respectively. The ordering of the markers within each chromosome was based on the recombination events between the markers. Linkage groups were compared with published tomato linkage maps.

QTL analysis was conducted using windows QTL Cartographer v 2.5 (Wang et al., 2010) software. The Composite Interval Mapping (CIM) method was used with the default parameters (model 6). A backward regression was used to perform the CIM analysis to enter or remove background markers from the model. The walking speed was set at one cM for the detection of QTL. The additive effect and the proportion of the phenotypic variation (R^2) for each QTL were also obtained using this software. A 1000 permutation option was chosen to determine the likelihood of an odd (LOD) score threshold to identify the presence of QTL in both environments (Li et al., 2007; Li et al., 2008; Meng et al., 2015). We used 5 cM scanning steps for the detection of QTL. The coefficient of variance (R^2-value), the relative contribution of genetic components, was calculated and described as the proportion of genetic variance explained by the QTL out of the total phenotypic variation. QTLs explaining more than 10% of the phenotypic variance were considered major QTLs, and QTLs found in at least two environments were considered to be consistent.

To designate each QTL, the letter '*q*,' followed by an abbreviation of EB resistance (EBR) was used as '*qEBR.*' Additionally, each QTL was classified by the chromosome in which a QTL was detected and then categorized by QTL number. Any QTL within a 5 cM distance on the same chromosome was regarded as a single QTL.

Results

Phenotypic data analysis

The disease symptoms of infected tomato plants in the greenhouse experiment varied from chlorotic and necrotic areas of leaves with concentric rings to defoliation and death. The two parental lines exhibited distinguished responses to EB, with NC 1CELBR being consistently resistant (disease score 3.0), and Fla. 7775 being susceptible (disease score 9.0) (Figure 1). The inoculated plants were scored for EB symptoms using a Horsfall-Barratt rating scheme (Horsfall and Barratt, 1945) at 14 and 21 days after inoculation. In field experiments, higher disease severities (6 to 11) were observed in 2011 (Figure 1A). There was a significant variation among $F_{2:3}$ lines for visual disease rating (Figure 1 and Table 1). Distribution of both field and greenhouse phenotypic data was continuous, indicating quantitative and polygenic control of EB resistance in tomatoes (Figure 1).

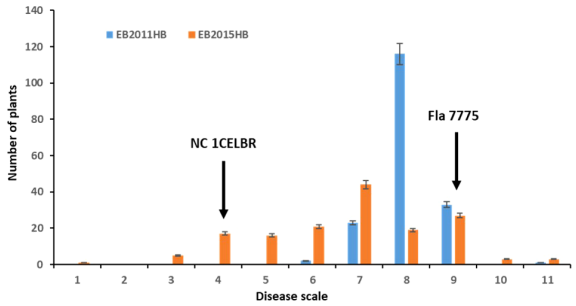


FIGURE 1
Frequency distribution for disease rating in a population of 174 F_2 and $F_{2:3}$ progenies. EB2011HB, the F_2 population was tested in a naturally-infected field at the Mountain Research Station, Waynesville, NC in 2011, and EB2015HB the $F_{2:3}$ progenies were evaluated in an artificial inoculation with a single *A. linariae* isolate in the greenhouse at Mountain Horticultural Research and Extension Center (MHCREC), Mills River, NC in 2015. Each inoculated plant was scored for EB symptoms using a Horsfall-Barratt rating scheme (Horsfall and Barratt, 1945). The values are the means of the parents and progenies, and arrows indicate resistant and susceptible parents. Bars denote the standard deviation.

The minimum and maximum EB development in 2011 in the population was 6 and 11, respectively, with an average of 8.1. In 2015, the minimum and maximum disease developments in this population were 1 and 11, with an average of 6.8 (Figure 1 and Table 1). These basic statistics over the years indicated that there was a good distribution of EB resistance in this population. The broad-sense heritability estimate for phenotypic data was 28.3%, and 25.3% for 2011, and 2015 disease evaluations, respectively. The disease score values showed a negative correlation between the years 2011 and 2015 (Figure 2A). The PCA bi-plot showed the possible association and high percentage of phenotypic variability was observed between the data sets of EB resistance in both environments (Figure 2B). The dimension of the first PC (Dim1) broadly outlined and explained 51.8% of the phenotypic variability for EB resistance in 2011 (Figure 1B). The dimension of the second PC (Dim2) also distinguished the 48.4% of phenotypic variability for EB resistance in 2015 at opposite angles of the PCA biplot (Figure 2B). This data also showed that EB resistance is controlled by multiple genes.

TABLE 1 Basic statistics of early blight development measured using a Horsfall and Barratt (1945) scale in the tomato population developed from NC 1CELBR × Fla. 7775.

Year	Environment	Sample size	Mean	Standard deviation	Minimum	Maximum	Variance	Heritability (%)
2011	Field (Waynesville)	174	8.1	0.62	6	11	0.36	28.3
2015	Greenhouse (Mills River)	174	6.8	1.7	1	11	9.61	25.3

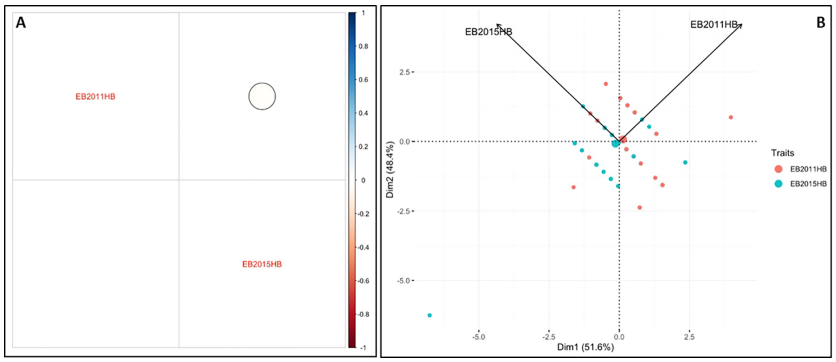


FIGURE 2
Analysis of phenotypic variability and correlation for early blight resistance in the mapping population. **(A)** Pearson's correlation between EB2011HB and EB2015HB **(B)** Principal component analysis (PCA) explains the potential phenotypic variability.

Linkage map construction of F_2

A total of 375 SNP markers were polymorphic between the parents. Those markers were used to genotype the population. A linkage map was constructed with these markers which covered approximately 737.17 cM genetic distance. The map results yielded a total of 12 linkage groups which are comparable with other tomato linkage maps and the number of tomato chromosomes. The Individual chromosomes had 18 to 65 markers with lengths ranging from 42.04 to 88.87 cM (Figure 3). Nearly 65 SNP markers were mapped on chromosome 4, followed by 42 SNP markers on chromosome 12 (Figure 3).

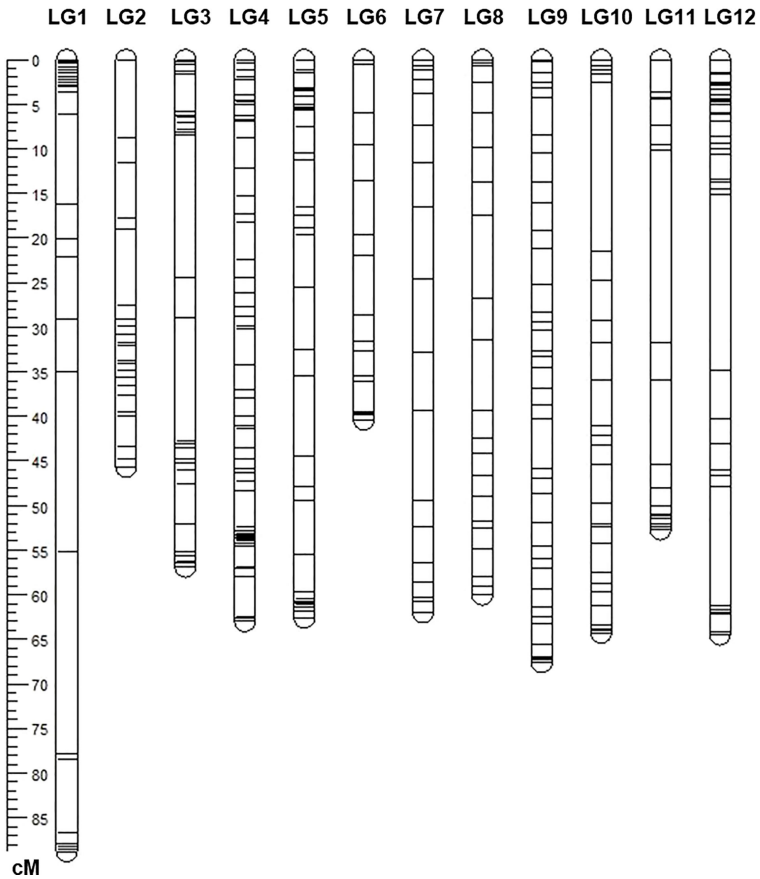


FIGURE 3
The linkage genetic map of the population of 174 F_2 progenies. The genetic map was developed from a cross between the resistant tomato line NC 1CELBR and the susceptible tomato cultivar Fla. 7775 using Solanaceae Coordinated Agricultural Project (SolCAP) derived Kompetitive Allele Specific PCR (KASP) markers.

QTL analysis

We identified QTLs for EB resistance using 174 $F_{2:3}$ derived lines and the SNP-based linkage map in two environments (Figure 4 and Table 2). In total, 6 QTLs, including major and minor effects, common for both environments were identified across the genome, explaining phenotypic variation (R^2) ranging from 3.8 to 21.0% (Figure 4 and Table 2). The QTLs on chromosomes 2, 8, and 11 (*qEBR2011-2*, *qEBR2011-8*, and *qEBR2011-11*) were detected in 2011, respectively. The QTLs *qEBR2011-2* (LOD: 4.2), *qEBR2011-8* (LOD: 4.2), and *qEBR2011-11* (LOD: 4.0) explained 3.8%, 12.1% and 11.7% of total phenotypic variations (Figure 4 and Table 2). The QTLs on same chromosomes were detected in 2015 as well (Figure 4 and Table 2). The QTLs *qEBR2015-2* (LOD: 5.0), *qEBR2015-8* (LOD: 5.2), and *qEBR2015-11* (LOD: 9.1) explained 21%, 11.4% and 19.8% of total phenotypic variations (Figure 4 and Table 2). We used the linked markers of the resistant QTLs to compare the resistance levels and allelic effects in the mapping population (Figure 5). As shown in the box plots, the homozygous resistant genotypes BB were associated with enhanced resistance compared to the homozygous susceptible genotype AA for all the QTLs in both environments (Figure 5). It also confirmed that all the resistant alleles in mapping population were inherited from NC 1CELBR. These results indicated that multiple genes/QTLs are contributing to EB resistance.

Discussion

We developed F_2 and F_2-derived mapping populations from a cross between the tomato breeding line NC 1CELBR (EB-resistant) and the susceptible tomato cultivar Fla. 7775 (EB-susceptible). The population was assessed for resistance to EB in the field trial and replicated greenhouse trials and genotyped with SNP molecular markers. Both field and greenhouse phenotypic data exhibited continuous distributions. The CIM analysis revealed 6 QTL conferring resistance to *A. linariae*. These QTLs explained up to 21% of the phenotypic variation confirming that genetic control for resistance to EB in NC 1CELBR is polygenic. The discovery of multiple QTL suggested that EB resistance in NC 1CELBR contributed different degrees of resistance to EB and behaved as a quantitatively inherited trait.

The estimate of broad-sense heritability (H^2) was 28.3% in the field test; whereas, in the greenhouse experiments it was 25.3%, suggesting a significant environmental effect on EB development in this mapping population. It is not surprising to have low narrow-sense heritability in this population since the heritability was determined from early (F_2 and F_3) generations. If the disease were evaluated at later generations, the level of homozygosity would go up, heterozygosity would go down, and resistance loci would have been fixed. The environmental effect could be minimized, and the genetic effect could be maximized, which is ultimately heritability. Disease severity was high in the 2011 field test, and presumably, this could be due to the dispersal of inoculum in the field, and within the plant canopy and variations in micro-climatic conditions, particularly dew and rain events, that would influence disease development during the tomato growing period (Rotem and Reichert, 1964). To avoid such confounding effects, phenotypic data are likely more reliable when large population sizes or even advanced populations such as recombinant-inbred lines (RILs) are evaluated in different environments with multiple replicates (Gardner, 1990). Nonetheless, we found the F_2 population had

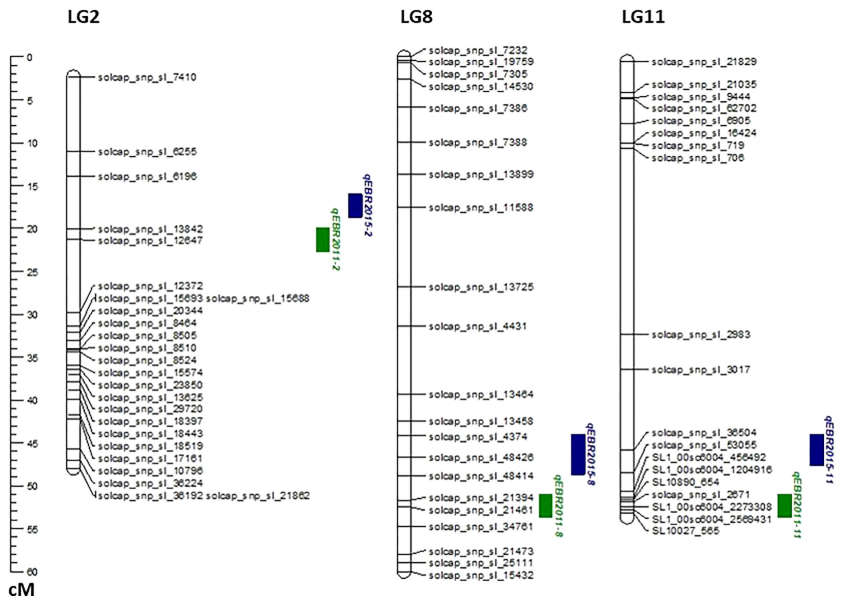


FIGURE 4
QTL analysis for early blight (EB) resistance in the $F_{2:3}$ mapping populations. Genetic linkage groups showing markers and the locations of EB-resistant QTLs in two different environments with the genetic distance shown in centimorgans (cM) for the mapping population evaluated during 2011and 2015.

10.3389/fpls.2023.1135884

TABLE 2 Quantitative trait loci (QTL) for early blight (EB) resistance in tomato detected by composite interval mapping (CIM) in a population of 174 $F_{2.3}$ progenies.

Trait	QTLs	Linkage group	Position (cM)	LOD	R^2 (%)	Additive	Dominant
EB2011HB	*qEBR2011-2*	2	20.01	4.17	3.8	-1.42	-2.53
EB2011HB	*qEBR2011-8*	8	51.31	4.18	12.1	-1.44	-2.64
EB2011HB	*qEBR2011-11*	11	50.91	4.03	11.7	-1.44	-2.65
EB2015HB	*qEBR2015-2*	2	16.61	5.02	21.0	0.71	-5.91
EB2015HB	*qEBR2015-8*	8	32.41	5.24	11.4	2.81	3.91
EB2015HB	*qEBR2015-11*	11	44.12	9.11	19.8	-2.19	-5.81

considerable resistance to EB and can be used to advance our effort to develop EB-resistant tomatoes and to combine multiple disease resistance with good fruit quality, which was started by releasing improved breeding lines and hybrids from our program before (Gardner and Panthee, 2010; Panthee and Gardner, 2010). Furthermore, NC 1 CELBR is the first identified tomato line that combines early blight resistance with the *Ph-2* and *Ph-3* genes for late blight resistance. The line was developed by performing crosses comprising wild species *S. habrochaites* and *S. pimpinellifolium* (Gardner and Panthee, 2010; Panthee and Gardner, 2010). It is worthwhile as parents in developing multiple disease resistant F_1 hybrids as well as parental lines for future tomato breeding programs with joint resistance to late blight and early blight without a linkage drag.

The results suggested that a functionally related QTLs conferring resistance to EB in the field and greenhouse had identical genetic regions. Although the QTLs were identified in the same genetic region, phenotypic variations in disease reaction between the field and greenhouse tests differed. In general, phenotypic variations in the 2011 field trial were lower compared to 2015 greenhouse trial. These results further emphasize that multiple replicated trials are necessary to conduct field EB evaluation and QTL identification. Furthermore, QTL detection is dependent on the level of precise phenotyping. We used foliar disease rating in the present study. Stem lesion was found to correlate better with the level of disease resistance, mainly when experiments are conducted in the greenhouse (Gardner, 1990). Anderson et al. (2021) have reported three QTLs from chromosomes 1, 5 and 9 based on foliar and stem lesions scoring. Therefore, it may be worth using stem lesions as well as foliar symptoms for EB QTL analysis in future studies.

Molecular markers and genetic maps are powerful tools to dissect complex traits and develop marker-assisted breeding strategies in tomatoes (Panthee and Chen, 2010; Foolad and Panthee, 2012). Foolad et al. (2002) developed BC_1, and BC_1S_1 populations of the *Solanum lycopersicum* x *S. habrachaites* cross and tested these in fields from 1998 - 2000. They identified ten major QTLs for resistance to EB using interval mapping. In another study, Zhang et al. (2003) identified six QTLs, four as major QTLs on chromosomes 5, 8, 10, and 11, and two as minor

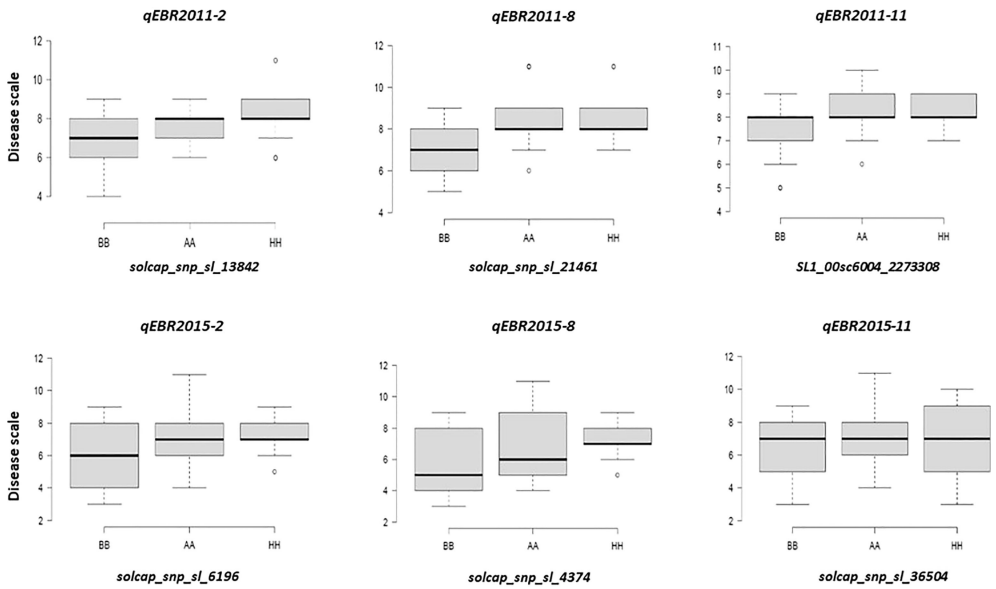


FIGURE 5
Box plots of resistance level regulated by linked markers to QTLs in F_2 segregating populations. Genotypes were grouped based on the associated SNP markers. AA: Fla. 7775, BB: NC 1CELBR, HH: Heterozygous.

QTLs on chromosomes 3 and 8. Both previous studies identified QTLs for resistance to EB using RFLP, SSR, and RGA markers (Foolad et al., 2002; Zhang et al., 2003), and they concluded that a high level of similarity between the two field studies was indicative of the stability of QTLs across populations and environments. In the present study, the reported QTLs were found in at least two experiments that were regarded as consistent QTLs as defined above. Although a different mapping population and markers were used, the QTLs detected on chromosomes 8 and 11 in this study agreed with the results of the previous studies (Foolad et al., 2002; Zhang et al., 2003). Ashrafi and Foolad (2015) identified four QTLs that are associated with EB from chromosomes 2, 5, 6, and 9. The positions of the QTLs found in the present study could not be compared because of the different marker types and genetic distance on the map. Furthermore, in the present study, even QTLs were detected at similar locations but the explained phenotypic variations were differ in different environments attributing to the environmental effects. The present study utilized SNP markers to identify QTLs resistance to EB and appeared to be useful for mapping and marker-assisted selection. Although we identified several SNP markers associated with QTLs for resistance to EB, these QTLs are likely to play distinct roles in plant defenses and plant innate immunity. The biological functions of these QTLs or genes in this pathosystem remain a critical unanswered question. Cloning, molecular characterization, and functional analysis of these QTLs in the tomato A. *linariae* interactions deserve further study.

Conclusion

The NC 1CELBR × Fla. 7775 derived mapping population was used to construct a genetic linkage map and QTL analysis for EB resistance. We detected a total of 6 QTLs, among them all QTLs conferring resistance to EB were inherited from NC 1CELBR. The SNP markers identified in this study are closely associated with putative EB- resistant QTLs and may be involved in host defense responses. To validate these results, additional mapping population development and fine mapping are necessary to determine their resistance spectrum to multiple isolates of *A. linariae*. Developing multiple advanced crosses and pyramiding resistance genes with superior quality is necessary to achieve enhanced resistance to early blight in tomatoes through MAS.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://zenodo.org/record/7677766#.Y_qoAXbMJRY.

Author contributions

TBA implemented the experiment and drafted the manuscript. MIS revised the manuscript, and analyzed the data. FJL revised the manuscript. SCS analyzed the data. DRP conceived the idea, designed the experiment, analyzed the data, and provided resources. All authors contributed to the article and approved the submitted version.

Funding

The Solanaceae Coordinated Agricultural Project (SolCAP) was funded by USDA/NIFA/AFRI grants (numbers 2008-55300-04757 and 2009-85606-05673) for developing SNP markers.

Acknowledgments

The technical assistance of Candice Anderson, Ragy Ibrahem, Adrienne Ratti, Tyler Nance, and Krishna Bhattarai in implementing the experiment is appreciated.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Adhikari, P., Oh, Y., and Panthee, D. R. (2017). Current status of early blight resistance in tomato: an update. *Int. J. Mol. Sci.* 18, 1–22. doi: 10.3390/ijms18102019

Amanullah, S., Osae, B. A., Yang, T., Abbas, F., Liu, S., Liu, H., et al. (2022). Mapping of genetic loci controlling fruit linked morphological traits of melon using developed CAPS markers. *Mol. Biol. Rep.* 49, 5459–5472. doi: 10.1007/s11033-022-07263-x

Anderson, T. A., Zitter, S. M., De Jong, D. M., Francis, D. M., and Mutschler, M. A. (2021). Cryptic introgressions contribute to transgressive segregation for early blight resistance in tomato. *Theor. Appl. Genet.* 134, 2561–2575. doi: 10.1007/s00122-021-03842-x

Asekova, S., Oh, E., Kulkarni, K. P., Siddique, M. I., Lee, M. H., Kim, J. I., et al. (2021). An integrated approach of QTL mapping and genome-wide association analysis identifies candidate genes for phytophthora blight resistance in sesame (*Sesamum indicum* l.). *Front. Plant Sci.* 12, 604709. doi: 10.3389/fpls.2021.604709

Ashrafi, H., and Foolad, M. R. (2015). Characterization of early blight resistance in a recombinant inbred line population of tomato: II. identification of QTLs and their co-localization with candidate resistance genes. *Adv. Stud. Biol.* 7, 149–168. doi: 10.12988/asb.2015.41163

Basu, P. (1974). Measuring early blight, its progress and influence on fruit losses in nine tomato cultivars. *Can. Plant Dis. Surv.* 54, 45–51.

Chaerani, R., Smulders, M. J. M., van der Linden, C. G., Vosman, B., Stam, P., and Voorrips, R. E. (2007). QTL identification for early blight resistance (*Alternaria solani*)

in a *Solanum lycopersicum* x s-arcanum cross. *Theor. Appl. Genet.* 114, 439–450. doi: 10.1007/s00122-006-0442-8

Cowgill, W. P., Maletta, M. H., Manning, T., Tietjen, W. H., Johnston, S. A., and Nitzsche, J. P. (2005). Early blight forecasting systems: evaluation, modification, and validation for use in fresh-market tomato production in northern new Jersey. *Hortscience* 40, 85–93. doi: 10.21273/HORTSCI.40.1.85

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics. 4th ed.* (London: Longman Group Ltd).

Foolad, M. R., and Panthee, D. R. (2012). Marker-assisted selection in tomato breeding. *Crit. Rev. Plant Sci.* 30, 93–123. doi: 10.1080/07352689.2011.616057

Foolad, M. R., Zhang, L. P., Khan, A. A., Nino-Liu, D., and Lin, G. Y. (2002). Identification of QTLs for early blight (*Alternaria solani*) resistance in tomato using backcross populations of a *Lycopersicon esculentum* × *l. hirsutum* cross. *Theor. Appl. Genet.* 104, 945–958. doi: 10.1007/s00122-002-0870-z

Gannibal, P. B., Orina, A. S., Mironenko, N. V., and Levitin, M. M. (2014). Differentiation of the closely related species, *Alternaria solani* and *A. tomatophila*, by molecular and morphological features and aggressiveness. *Eur. J. Plant Pathol.* 139, 609–623. doi: 10.1007/s10658-014-0417-6

Gardner, R. G. (1984). Use of *Lycopersicon hirsutum* PI 126445 in breeding early blight-resistant tomatoes. *Hortscience* 19, 208–208.

Gardner, R. G. (1988). NC-EBR-1 and NC-EBR-2 early blight resistant tomato breeding lines. *Hortscience* 23, 779–781. doi: 10.21273/HORTSCI.23.4.779

Gardner, R. G. (1990). Greenhouse disease screen facilitates breeding resistance to tomato early blight. *HortScience* 25, 222–223. doi: 10.21273/HORTSCI.25.2.222

Gardner, R. G., and Panthee, D. R. (2010). NC 1 CELBR and NC 2 CELBR: early blight and late blight resistant fresh market tomato breeding lines. *HortScience* 45, 975–976. doi: 10.21273/HORTSCI.45.6.975

Gardner, R. G., and Panthee, D. R. (2012). 'Mountain magic': an early blight and late blight-resistant specialty type f-1 hybrid tomato. *Hortscience* 47, 299–300. doi: 10.21273/HORTSCI.47.2.299

Gardner, R. G., and Shoemaker, P. B. (1999). 'Mountain supreme' early blight-resistant hybrid tomato and its parents, NC EBR-3 and NC EBR-4. *Hortscience* 34, 745–746. doi: 10.21273/HORTSCI.34.4.745

Gleason, M. L., Macnab, A. A., Pitblado, R. E., Ricker, M. D., East, D. A., and Latin, R. X. (1995). Disease-warning systems for processing tomatoes in eastern north America – are we there yet. *Plant Dis.* 79, 113–121. doi: 10.1094/PD-79-0113

Horsfall, J. G., and Barratt, R. W. (1945). An improved grading system for measuring plant diseases. *Phytopathology* 35, 655–655.

Ivors, K. (2011). 2011 foliar fungicide spray guide for tomatoes in NC 2013.

Ivors, K. L., Mill, D. C., and Holmberg, C. (2007). Evaluation of spray programs for control of early blight and late blight of tomato, 2007. *Plant Dis. Manage. Rep.* 2, V096.

Jiménez-Gómez, J. M., and Maloof, J. N. (2009). Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. *BMC Plant Biol.* 9, 85. doi: 10.1186/1471-2229-9-85

Jones, J. P. (1991). "Early blight," in *Compendium of tomato diseases*. Eds. J. B. Jones, J. P. Jones, R. E. Stall and T. A. Zitter (St. Paul, MN: APS Press).

Keinath, A. P., DuBose, V. B., and Rathwell, P. J. (1996). Efficacy and economics of three fungicide application schedules for early blight control and yield of fresh-market tomato. *Plant Dis.* 80, 1277–1282. doi: 10.1094/PD-80-1277

Kosambi, D. D. (1943). The estimation of map distances from recombination values. *Ann. Eugen.* 12, 172–175. doi: 10.1111/j.1469-1809.1943.tb02321.x

Li, H. H., Ribaut, J. M., Li, Z. L., and Wang, J. K. (2008). Inclusive composite interval mapping (ICIM) for digenic epistasis of quantitative traits in biparental populations. *Theor. Appl. Genet.* 116, 243–260. doi: 10.1007/s00122-007-0663-5

Li, H. H., Ye, G. Y., and Wang, J. K. (2007). A modified algorithm for the improvement of composite interval mapping. *Genetics* 175, 361–374. doi: 10.1534/genetics.106.066811

Louws, F. J., Hausbeck, M. K., Kelly, J. F., and Stephens, C. T. (1996). Impact of reduced fungicide and tillage on foliar blight, fruit rot, and yield of processing tomatoes. *Plant Dis.* 80, 1251–1256. doi: 10.1094/PD-80-1251

Madden, L., Pennypacker, S. P., and Macnab, A. A. (1978). FAST, a forecast system for *Alternaria solani* on tomato. *Phytopathology* 68, 1354–1358. doi: 10.1094/Phyto-68-1354

Meng, L., Li, H. H., Zhang, L. Y., and Wang, J. K. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Nash, A. F., and Gardner, R. G. (1988). Tomato early blight resistance in a breeding line derived from *Lycopersicon hirsutum* PI 126445. *Plant Dis.* 72, 206–209. doi: 10.1094/PD-72-0206

Nyquist, W. E. (1991). Estimation of heritability and prediction of selection response in plant populations. *Crit. Rev. Plant Sci.* 10, 235–322. doi: 10.1080/07352689109382313

Ohlson, E. W., and Foolad, M. R. (2015). Heritability of late blight resistance in tomato conferred by *Solanum pimpinellifolium* accession PI 224710. *Plant Breed.* 134, 461–467. doi: 10.1111/pbr.12273

Panthee, D. R., and Chen, F. (2010). Genomics of fungal disease resistance in tomato. *Curr. Genomics* 11, 30–39. doi: 10.2174/138920210790217927

Panthee, D. R., and Gardner, R. G. (2010). 'Mountain merit': a late blight-resistant large-fruited tomato hybrid. *HortScience* 45, 1547–1548. doi: 10.21273/HORTSCI.45.10.1547

Pasche, J. S., and Gudmestad, N. C. (2008). Prevalence, competitive fitness and impact of the F129L mutation in *Alternaria solani* from the united states. *Crop Prot.* 27, 427–435. doi: 10.1016/j.cropro.2007.07.011

Pasche, J. S., Piche, L. M., and Gudmestad, N. C. (2005). Effect of the F129L mutation in *Alternaria solani* on fungicides affecting mitochondrial respiration. *Plant Dis.* 89, 269–278. doi: 10.1094/PD-89-0269

Pennypacker, S. P., Madden, L. V., and Macnab, A. A. (1983). Validation of an early blight forecasting system for tomatoes. *Plant Dis.* 67, 287–289. doi: 10.1094/PD-67-287

Pitblado, R. (1992). *The development and implementation of TOM-CAST, a weather-timed fungicide spray program for field tomatoes* (Ridgetown, Ontario, Canada: Ministry of Agricultural Technology).

R Core Team (2018). *R: a language and environment for statistical computing* (Vienna: R foundation for statistical computing).

Rotem, J. (1994). *The genus alternaria: biology, epidemiology, and pathogenicity* (St. Paul, Minnesota: American Phytopathological Society).

Rotem, J., and Reichert, I. (1964). Dew - a principal moisture factor enabling early blight epidemics in a semiarid region of Israel. *Plant Dis. Rep.* 48, 211–221.

SAS Institute Inc. (2012). *The SAS system, version 9.4 for windows. 9th ed* (Cary, NC: SAS Institute).

Sim, S.-C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., et al. (2012a). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PloS One* 7, e40563. doi: 10.1371/journal.pone.0040563

Sim, S. C., Van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., et al. (2012b). High-density SNP genotyping of tomato (*Solanum lycopersicum* l.) reveals patterns of genetic variation due to breeding. *PloS One* 7, e45520. doi: 10.1371/journal.pone.0045520

van Ooijen, J. W. (2006). *Joinmap 4.0: software for the calculation of genetic linkage maps in experimental populations* (Wageningen: Kyazma B.V).

Wang, S. C., Basten, J., Gaffney, P., and Zeng, Z. B. (2010). *Windows QTL cartographer department of statistics* (Raleigh, NC: North Carolina State University).

Wurschum, T. (2012). Mapping QTL for agronomic traits in breeding populations. *Theor. Appl. Genet.* 125, 201–210. doi: 10.1007/s00122-012-1887-6

Zhang, L. P., Lin, G. Y., Nino-Liu, D., and Foolad, M. R. (2003). Mapping QTLs conferring early blight (*Alternaria solani*) resistance in a *Lycopersicon esculentum* × *l. hirsutum* cross by selective genotyping. *Mol. Breed* 12, 3–19. doi: 10.1023/A:1025434319940

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fpls.2023.1201103

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Pei Xu,
China Jiliang University, China
Luming Yang,
Henan Agricultural University, China

*CORRESPONDENCE
Libin Wei
libinwei2013@aliyun.com
Xuejun Wang
wangxj4002@sina.com

RECEIVED 06 April 2023
ACCEPTED 10 May 2023
PUBLISHED 07 June 2023

CITATION
Zhao N, Xue D, Miao Y, Wang Y, Zhou E, Zhou Y, Yao M, Gu C, Wang K, Li B, Wei L and Wang X (2023) Construction of a high-density genetic map for faba bean (*Vicia faba* L.) and quantitative trait loci mapping of seed-related traits.
Front. Plant Sci. 14:1201103.
doi: 10.3389/fpls.2023.1201103

Construction of a high-density genetic map for faba bean (*Vicia faba* L.) and quantitative trait loci mapping of seed-related traits

Na Zhao, Dong Xue, Yamei Miao, Yongqiang Wang, Enqiang Zhou, Yao Zhou, Mengnan Yao, Chunyan Gu, Kaihua Wang, Bo Li, Libin Wei* and Xuejun Wang*

Department of Economic Crops, Jiangsu Yanjiang Institute of Agricultural Science, Nantong, China

Faba bean (*Vicia faba* L.) is a valuable legume crop and data on its seed-related traits is required for yield and quality improvements. However, basic research on faba bean is lagging compared to that of other major crops. In this study, an F_2 faba bean population, including 121 plants derived from the cross WY7×TCX7, was genotyped using the Faba_bean_130 K targeted next-generation sequencing genotyping platform. The data were used to construct the first ultra-dense faba bean genetic map consisting of 12,023 single nucleotide polymorphisms markers covering 1,182.65 cM with an average distance of 0.098 cM. The map consisted of 6 linkage groups, which is consistent with the 6 faba bean chromosome pairs. A total of 65 quantitative trait loci (QTL) for seed-related traits were identified (3 for 100-seed weight, 28 for seed shape, 12 for seed coat color, and 22 for nutritional quality). Furthermore, 333 candidate genes that are likely to participate in the regulation of seed-related traits were also identified. Our research findings can provide a basis for future faba bean marker-assisted breeding and be helpful to further modify and improve the reference genome.

KEYWORDS

***vicia faba* L., single nucleotide polymorphisms (SNP), high-density genetic map, seed related traits, quantitative trait loci (QTL)**

1 Introduction

Faba bean (*Vicia faba* L.), also called horse bean, is a member of the Fabaceae family (grain legume) that originated in the Near East, and is an important cool-season food legume (Cubero, 1974). It is currently widely cultivated in Africa, Asia, Europe, Australia, and North America (Alghamdi et al., 2012). Faba bean can be used as a green manure as it has nitrogen fixation capabilities and can thus improve soil quality (Jensen et al., 2010).

Additionally, faba bean is used as a type of food for humans and as a feed for animals (Martineau-Côté et al., 2022), and the fresh seeds can be consumed as vegetables (Zong et al., 2009). Furthermore, due to its rich nutritional value and high protein and lysine content, it can be effectively utilized as a source of plant protein (Etemadia et al., 2019), and is also rich in phenols (Amarowicz and Shahidi, 2017). The edible part of the seed thus directly affects its yield and quality. It is therefore important to clarify the genetic basis of the related traits in faba bean breeding programs. The phenotypic and quality traits of faba bean seeds are mostly complex and easily affected by the environment, and consequently, the use of molecular technologies is required to fully understand them. The genetic linkage map is an effective tool that can help to improve our understanding of the inheritance of traits at a genome-wide level (Verma et al., 2015). Furthermore, the fine mapping of quantitative trait loci (QTL) and candidate genes related to specific traits has traditionally been performed using high-resolution genetic linkage maps (Zhang et al., 2016).

Faba bean has one of the largest genomes among crop legumes, and is diploid with 2n = 12 chromosomes and a large genome of 13,000 Mb (Johnston et al., 1999). As a result, basic research on faba bean is lagging behind that of other major crops that have relatively complete genetic maps, such as maize (*Zea mays* L.), rice (*Oryza sativa* L.), and wheat (*Triticum aestivum* L.) (Wang H et al., 20122). Initially, some traditional markers, including morphological and isoenzyme, random amplified polymorphic DNA, and microsatellite markers, were used to construct several faba bean genetic maps (Torres et al., 1993; Satovic et al., 1996; Patto et al., 1999; Román et al., 2002; Ávila et al., 2004; Román et al., 2004; Ávila et al., 2005; Ellwood et al., 2008; Díaz-Ruiz et al., 2009; Díaz-Ruiz et al., 2010; Cruz-Izquierdo et al., 2012; Gutiérrez et al., 2013). With the development of high-throughput sequencing technologies, simple sequence repeats (SSR) and single nucleotide polymorphisms (SNP) have been extensively used to construct genetic maps and identify QTLs in faba beans (Arbaoui et al., 2008; Ma et al., 2013; Satovic et al., 2013; Kaur et al., 2014; Sallam et al., 2016; Webb et al., 2016; Catt et al., 2017; Ocaña–Moral et al., 2017; Yang et al., 2019). Sudheesh et al. (2019) constructed an integrated genetic map for faba bean spanning 1,439 cM, with an average distance of 0.80 cM per marker using a total of 1,850 markers. Carrillo-perdomo et al. (2020) constructed a high-density genetic map containing gene-based SNP markers with a length of 1,547.71 cM, and an average distance of 0.89 cM. Recently, an integrated genetic linkage map containing 6,895 SNPs, with a length of 3,324.48 cM was constructed from two F_2 populations by Li et al. (2023).The construction of a fine linkage map for faba bean can greatly improve the efficiency of related genetic research and crop breeding and enable the establishment of marker selection and QTL mapping associated with economically important traits (Khazaei et al., 2014; Aguilar–Benitez et al., 2021; Gutierrez and Torres, 2021; Carrillo-Perdomo et al., 2022).

To date, although there have been some studies on QTL mapping associated with seed-related traits in faba bean, few related genes have been mapped. The QTL associated with 100-seed weight was first identified on chromosome 6 and significantly correlated with 20 markers (Patto et al., 1999). Furthermore, Ávila et al. (2017) identified five QTLs related to 100-seed weight. The F_2 populations generated from Yun122 and TF42 were used to construct genetic maps, and four QTLs controlling seed length, width, and 100-seed weight were identified (Tian et al., 2018). Macas et al. 1993a mapped the chromosomal positions of genes encoding seed storage proteins. Gutierrez et al. (2007) identified two SCAR markers tightly linked to a gene controlling tannin deficiency in faba beans and Hou et al. (2018) screened one SSR marker (SSR84) closely linked to the tannin content (zt-1) gene using 596 SSR markers and 100 ISSR markers, which could aid in accurate prediction of the zt-1 genotypes. Recently, 15 markers were identified with seed size associations based on genome-wide association study (Jayakodi et al., 2023). Li et al. (2023) identified 32 QTLs related to seed size and 6 QTLs related to seed coat color.

The efficiency and precision of QTL mapping are restricted by the low density of molecular markers in the resulting genetic maps; however, this can be addressed using high-throughput DNA microarray (DNA chip) technologies. Wang et al. (2021) utilized a large-scale transcriptome and a large number of SNP markers to develop the Faba_bean_ 130 K SNP targeted next-generation sequencing (TNGS) genotyping platform, which contains 130,514 SNPs and can be used for high-density genetic linkage map development and QTL mapping.

In this study, an ultra-dense genetic map from an F_2 population was constructed using the Faba_bean_ 130 K SNP TNGS genotyping platform. QTLs for 15 seed-related traits, including 100-seed weight (HSW), seed area (SA), seed perimeter (SP), seed length (SL), seed width (SW), seed length and width ratio (SLWR), seed thickness (ST), seed coat color R value (SC-R), seed coat color G value (SC-G), seed coat color B value (SC-B), protein content (PC), starch content (StC), fiber content (FC), lipid content (LC), and tannin content (TC), which were mapped based on the phenotypic data from F_2 and $F_{2:3}$ populations. The ultra-dense genetic map and QTLs produced from this study can be used for faba bean marker-assisted selection (MAS), gene mapping, and reference genome improving. MAS is a method used in plant breeding, once the linkage has been established between physical markers and the target traits, individuals with desirable traits can be selected by detecting the molecular markers.

2 Article types

This article was submitted to Plant Breeding, a section of the journal Frontiers in Plant Science.

Abbreviations: CV, Coefficient of variation; FC, Fiber content; HSW, 100-seed weight; LG, Linkage group; LC, Lipid content; LOD, Logarithm of odds; MAS, Marker-assisted selection; PC, Protein content; QTL, Quantitative trait loci; SA, Seed area; SC-B, Seed coat color B value; SC-G, Seed coat color G value; SC-R, Seed coat color R value; SD, Standard deviation; SL, Seed length; SLWR Seed length and width ratio; SNP, Single-nucleotide polymorphism; SP, Seed perimeter; SSR, Simple sequence repeats; ST, Seed thickness; StC, Starch content; SW, Seed width; TC, Tannin content; TNGS, Targeted next-generation sequencing.

3 Materials and methods

3.1 Plant materials and phenotypic data evaluation

An interspecific F_2 population containing 121 individual plants was generated from WY7 and TCX7 parent materials. The female parent WY7 is a germplasm resource introduced from the UK with a medium seed size and dark-purple seed coat color. The male parent TCX7 has a large seed size, with a white seed coat, is of good quality, and is cultivated by the Jiangsu Yanjiang Institute of Agricultural Sciences, China. The F_2 individual plants and their parents were planted in Xueyao, Jiangsu Province, China from 2020–2021, and the $F_{2:3}$ plant lines and their parents were planted in Xueyao and Jiuhua respectively, Jiangsu Province, China, from 2021–2022. Each faba bean line was planted in one row of 2.4 m in length, with a row distance of 0.8 m, and plant spacing of 0.2 m. Field management was consistent with local production practices throughout the whole growth period. Ten seed phenotypic traits and five nutritional quality traits of the parents, F_2 individual plants, and $F_{2:3}$ families in two environments (Xueyao and Jiuhua) were investigated. Ten plants in each $F_{2:3}$ line and their parents were harvested. The seed shape traits assessed were SA, SP, SL, SW, SLWR, and ST. The average indicators of HSW, SA, SP, SL, SW, and SLWR used an automatic seed testing system (SC-A1, Hangzhou Wanshen Detection Technology Co., Ltd., Hangzhou, China). The average values of the 10 thickest parts of the seeds were regarded as ST. The seed coat color traits including SC-R, SC-G, and SC-B were measured by spectrophotometer (YS3020, 3NH, China). Nutritional quality traits PC, StC, FC, LC, and TC were determined using a DA7250 NIR analyzer (Perten Instruments, Hägersten, Sweden) with three replicates.

Statistical analysis of the data, such as frequency distribution, coefficient of variation, standard deviation, skewness and kurtosis analysis, was performed using the ANOVA function of IciMapping 4.2.53. The phenotypic correlation between these traits was obtained by Pearon's correlation analyses using SPSS software. Ver. 26 (IBM SPSS Statistics, Chicago, IL, USA) and R software (version 3.2.2, http://www.r-project.org).

3.2 Genotyping

The total genomic DNA of the F_2 individuals and their parental lines was extracted from fresh leaves using the CTAB method (Doyle and Doyle, 1987). A NanoDrop spectrophotometer (Thermo Fisher Scientific, USA) was used to determine the optical density ratios of OD260/280 (>1.8) and OD260/230 >1.5). A Qubit was used for precise quantification, and gel electrophoresis was used to monitor and assess the quality and contamination of all DNA samples. The Illumina sequencing library was constructed by binding biotin-labeled RNA probes to spliced DNA fragments using restriction enzymes and was sequenced using the China Golden Maker (Beijing) Biotech Co. Clean data were derived from the raw sequencing data after quality control (filter parameters: trimmomatic-0.36.jar PE -phred33 ILLUMINACLIP: fa: 2:30:10:8: true LEADING:3 TRAILING:3 SLIDINGWINDOW: 4:15 MINLEN:100) and then matched to the faba bean transcriptome (Wang et al., 2021) by using BWA software (version 0.7.17) with parameters: MEM -T 4 -K 32 -M -R). Based on the results of the sequence alignment, SNPs from the populations genomic data were detected with GATK (version 4.1.2.0) and filtered with VCFtools (version 0.1.13). The detailed criteria and analysis methods were in accordance with Wang et al. (2021).

3.3 Construction of the genetic map

The harvested genotypes of the samples were firstly filtered before genetic map construction. Based on the filtered genotypes, for each loci, the individuals were coded as "A" (if same with parent TCX7), "B" (same with the parent WY7), "H" (heterozygous containing 2 alleles from each of the parents) or "missing" (all other scenarios). The discarded loci include 1) the loci which were heterozygous in either parent, and 2) the loci with the missing rate above 80% in the population. This was done by using a python script from Li et al. (2021). The genetic map construction used a similar procedure as Li et al. (2021). Briefly, the coded "ABH" genotype matrix was firstly filtered to discarding distortion loci with the threshold P value = 0.01, and then was fed to Lep-Map3 (Rastas, 2017). The default parameters and a logarithm of odds (LOD) score of 12 were used in Lep-Map3. Linkage groups (LGs) with markers less than 100 were removed, and Kosambi function was applied to covert the recombinant rate into LG length (cM, centi-Morgan).

3.4 QTL mapping

Based on the genetic map constructed above and the phenotypes from multiple environments, we conducted QTL mapping in 2 programs, i.e. QTL Cartographer 2.5 (Wang S et al., 2012) and IciMapping 4.2.53 (Meng et al., 2015). In QTL Cartographer, CIM (Composite interval mapping) method was used, and the parameters were set up as: control markers = 5, window size = 10.0 cM, walk speed = 1.0 cM, and the LOD threshold was determined by 500 times permutation tests. For the Icimapping program, ICIM (Inclusive composite interval mapping) method was selected, and the flowing parameters were used: "missing phenotype = Deletion", "mapping step = 1 cM" and "LOD threshold = 1000 times permutation at type I error 0.05". In the mapping result, VG/VP value reflects the explanation rate of phenotypic variance, and the confidence interval of a QTL was determined by the outermost 2 markers above threshold. The QTLs were named as follows: q + trait abbreviation + chromosome number + QTL number.

3.5 Candidate gene identification and annotation

The splice junction sequences in the Faba_bean_ 130 K SNP TNGS genotyping platform were searched within the QTL intervals and then mapped to 243,120 unigenes (Wang et al., 2021), which were referred to in order to obtain the candidate genes and their gene annotations.

3.6 Reference genome mapping

Sequence of the genes in genetic map alignment with reference genome (https://projects.au.dk/fabagenome/genomics-data) and the candidate genes were visual mapped to the reference genome using TB tools software.

4 Results

4.1 Phenotypic analyses

The two parent materials showed significant differences in HSW, SA, SP, SL, SW, SC-R, SC-G, SC-B, FC, TC, StC and LC (Table 1). The statistical results of the phenotypic variations in the seed-related traits among the parents, F_2 populations, and $F_{2:3}$ individuals (Supplementary Table S1) suggested that HSW, SA, SP, SL, SW, SLWR, PC, StC, FC, LC, and TC showed continuous variation. The absolute values of skewness and kurtosis were almost less than 1, approximately conforming to the normal distribution, meeting the requirements of QTL analysis (Figure 1; Table 1).

4.2 Correlation analyses among different traits

Significant Pearson's correlations ($p < 0.01$) for the same trait showed a significant positive relationship between the F_2 and $F_{2:3}$ populations in Xueyao and Jiuhua (Supplementary Table S2). Phenotypic correlations ($p < 0.01$) among the different traits are shown in Figure 2. Seed shape traits, including HSW, SA, SP, SL, and SW, were positively correlated with PC, and negatively correlated with StC. Seed coat color traits, including SC-R, SC-G, and SC-B, were positively correlated with FC and StC and negatively correlated with LC. There was no significant correlation between the seed shape traits and seed coat color traits in this study.

4.3 Genetic map construction

A total of 121 F_2 plants and their parents were genotyped using 130,514 SNPs in the Faba_bean_ 130 K SNP TNGS genotyping platform, showing excellent results, quality, and matching scores (Supplementary Tables S3; S4). There were 12,023 SNP-tagged gene microarrays with polymorphism between parents (Supplementary

TABLE 1 Details of average of F_2, two environments of $F_{2:3}$ individuals and their parents.

Trait	Parents		Population							
	WY7	TCX7	Min	Max	Mean	SD	Variance	CV%	Skewness	Kurtosis
HSW	133.05**	245.54**	113.62	230.34	170.20	27.72	762.07	16.29	0.23	-0.69
SA	290.20 **	526.83 **	245.00	484.25	360.25	59.28	3,485.39	16.46	0.24	-0.85
SP	65.46**	89.15 **	55.94	87.59	73.38	6.53	42.26	8.90	0.08	-0.59
SL	22.56 **	30.61 **	19.18	29.60	25.22	2.13	4.52	8.46	0.00	-0.46
SW	16.00**	21.62**	13.56	20.88	17.63	1.55	2.37	8.77	0.09	-0.65
SLWR	1.43	1.42	1.30	1.53	1.44	0.05	0.00	3.46	-0.33	-0.35
ST	9.03	9.97	6.67	10.92	8.93	0.71	0.49	7.91	-0.03	0.28
SC-R	64.22**	136.62 **	53.34	158.93	94.16	31.80	1,002.93	33.77	0.64	-1.12
SC-G	56.89*	119.36 *	46.84	137.25	80.35	25.73	656.80	32.03	0.70	-0.98
SC-B	61.94 *	91.37 *	54.26	104.06	74.90	11.92	141.00	15.92	0.49	-0.70
FC	5.67 *	10.95*	3.02	12.83	7.98	2.06	4.19	25.78	-0.15	-0.35
TC	0.53 **	0.65 **	0.43	0.72	0.57	0.06	0.00	9.84	0.16	-0.27
StC	32.66 *	35.95 *	29.56	37.58	33.95	1.53	2.32	4.50	-0.23	0.42
PC	30.90	31.40	28.19	34.05	31.12	1.22	1.48	3.93	-0.19	-0.29
LC	1.30 *	1.01 *	0.83	1.48	1.12	0.13	0.02	11.53	0.07	-0.23

SD standard deviation, CV coefficient of variation, HSW 100-seed weight (g), SA seed surface area (mm^2), SP seed perimeter (mm), SL seed length (mm), SW seed width (mm), SLWR seed length and width ratio, ST seed thickness (mm), SC-R seed coat color R value, SC-G seed coat color G value, SC-B seed coat color B value, FC fiber content (%), TC tannin content (%), StC starch content (%), PC protein content (%), LC lipid content (%). Significant differences between two parental lines WY7 and TCX7 are marked by * and **, which were determined by the Student's t test at $P < 0.05$ and $P < 0.01$, respectively.

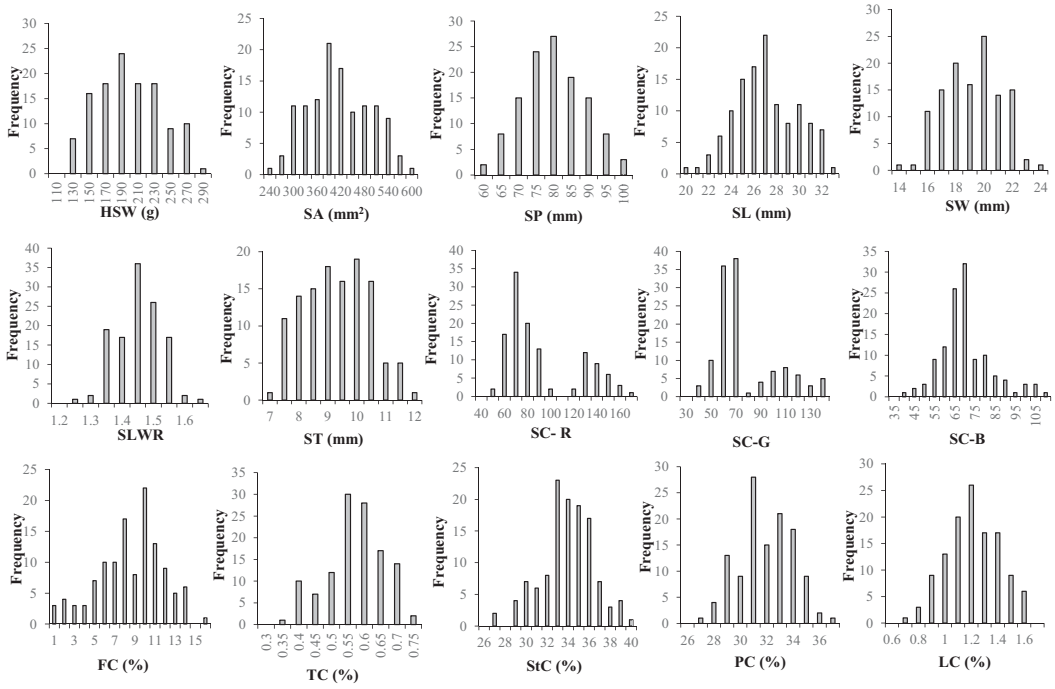


FIGURE 1
Frequency distributions of seed-related traits in 121 F_2 derived from a cross between WY7 andTCX7. HSW 100-seed weight (g), SA seed surface area (mm^2), SP seed perimeter (mm), SL seed length (mm), SW seed width (mm), SLWR seed length and width ratio, ST seed thickness (mm), SC-R seed coat color R value, SC-G seed coat color G value, SC-B seed coat color B value, FC fiber content (%), TC tannin content (%), StC starch content (%), PC protein content (%), LC lipid content (%).

Table S5), and they were successfully genotyped into "A," "B," and "H" types in the population. All co-isolated markers were defined as one bin, and 1106 bin markers were used to construct a genetic map containing 6 LGs. The overall length of the genetic map was 1,182.65 cM with an average marker spacing of 0.098 cM. Each LG range was from 157.08–296.82 cM, and the average distance between markers was from 0.079–0.114 cM. LG1 had the largest number of markers with 3,325 SNPs. The smallest gap identified in the map was 0.826 cM, the total number of gaps > 5 cM was 9, and the largest gap was 11.78 cM LG6. Additionally, the ratio of marker intervals < 5 cM for all LGs was > 97% (Figure 3; Table 2).

4.4 QTL analysis

QTL mapping was performed using QTL lciMapping and QTL-Cart CIM, and 65 (Supplementary Table S6) and 50 (Supplementary Table S7) QTLs were identified for all 15 seed-related traits detected in the F_2 and $F_{2:3}$ populations, respectively. Together, these two mapping strategies identified 28 overlapping QTLs (Supplementary Table S8). Of these, the QTL intervals observed using the CIM method were usually wider, whereas the intervals from the ICIM method were narrower. Consequently, the results obtained using the lCIM method were used in this study. The genetic effect (the explanation rate of phenotyte variance or VG/VP) of the QTLs detected using ICIM for 15 seed-related traits ranged from 4.90–73.99%, with peak LOD values ranging from 4.48–35.25 (Supplementary Table S6). Among the 65 loci, there were 11 QTLs that were detected for more than two traits (Supplementary Table S9). There were 39 QTLs identified that individually accounted for > 10% of the phenotypic variation (Table 3) and 1 QTL explained < 5% of the phenotypic variation (Supplementary Table S6). A total of 41 and 21 QTLs were found to have positive and negative additive effects, respectively.

4.4.1 Seed morphology traits

Three QTLs of HSW were detected and had peak LOD scores of 4.64–17.81, which explained 7.26%–38.88% of the HSW variation. One was located on LG4, and two were mapped to LG5 (Table 3). QTLs detected more than two times among F_2, F_3-XY and F_3-JH were considered environmentally stable. *qHSW5.1* was detected in F_2 and F_3-JH (Table 3; Supplementary Table S10).

A total of 28 QTLs were found for several seed shape traits, and 6 were regarded as stable (Table 3; Supplementary Table S10). Five QTLs were detected on LG5 with a peak LOD score of 4.69–24.81, and they explained 4.90–51.51% of the SA variation. *qSA5.1* was detected in F_2 and F_3-XY. Only one environmentally stable QTL (*qSP5.1*) of SP was identified on LG5 with a peak LOD score ranging from 14.82–23.55, and it explained 40.13–55.47% of the SP variation. Four QTLs associated with SL had peak LOD scores ranging 4.73–28.30, which explained 5.50–48.32% of the SL variation and were located on LG1, LG3, LG5, and LG6. According to the results, *qSL5.1* was a stable QTL, which detected in F_2, F_3-XY and F_3-JH.

10.3389/fpls.2023.1201103

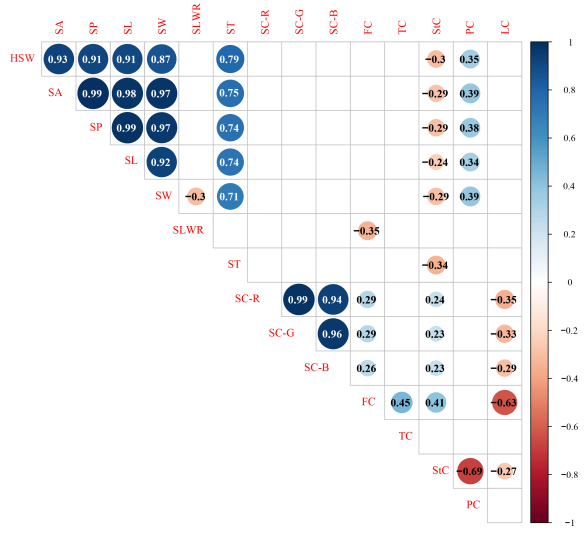


FIGURE 2
Correlation analysis of different traits in 0.01 probability level.HSW 100-seed weight (g), SA seed surface area (mm^2), SP seed perimeter (mm), SL seed length (mm), SW seed width (mm), SLWR seed length and width ratio, ST seed thickness (mm), SC-R seed coat color R value, SC-G seed coat color G value, SC-B seed coat color B value, FC fiber content (%), TC tannin content (%), StC starch content (%), PC protein content (%), LC lipid content (%).

Five QTLs explained 6.64–42.93% of the SW variance, with peak LOD scores ranging from 4.48–22.55, which were identified in linkage groups LG2 (1), LG3 (1), LG5 (1), and LG6 (2). An environmentally stable QTL (*qSW5.1*) was also identified. Five QTLs explained 9.59–25.02% of the SLWR variance, and the peak LOD scores varied from 5.33–11.11, and *qSW6.1* was stable.

For ST, eight QTLs were detected in LG4 (3), LG5 (3), and LG6 (2), with LOD scores ranging from 4.48–12.25, and they explained 6.65–24.78% of the total phenotypic variation. *qST5.1* was detected in F_2, F_3-XY and F_3-JH. Among these QTLs, four were overlapping for seed shape traits.

For seed coat color traits, 12 QTLs were detected, including 3, 5, and 4 QTLs for R, G, and B, respectively. The phenotypic variation explained by each individual QTL ranged from 5.00–73.99%, with a peak LOD of 4.53–35.25 (Table 3). Three were overlapping QTLs and one was a stable QTL, both located in linkage group LG1 (Supplementary Table S10).

4.4.2 Nutritional quality traits

The results from the QTL analysis identified 22 QTLs associated with nutritional quality traits (Table 3; Supplementary Table S10), 7 QTLs explained 9.25–21.35% of the FC variance, 2 QTLs explained 22.70–17.61% of the TC variance, 7 QTLs explained 7.09–18.44% of the StC variance, 2 QTLs explained 13.65–17.32% of the PC variance, and 4 QTLs explained 6.74–21.35% of LC variance. *qFC3. 3* was considered stable.

4.5 Analysis of candidate genes

The genes in the QTL intervals were screened using the Faba_bean_ 130 K SNP TNGS genotyping platform (Table 4). The results showed that 333 genes and 610 SNPs were detected at 65 QTL intervals. Among the 333 genes, HSW, seed shape, seed coat color, and nutritional quality traits contained 8, 117, 100, and 109 genes, respectively, and 173 genes were functionally annotated by database comparison. A total of 67 candidate genes within the environmentally stable QTL intervals were detected, including 2, 20, 39, 3 genes related to HSW, seed shape, seed coat color, and nutritional quality traits, respectively. The results showed that 213 genes in 41 QTLs explained > 10% of the observed phenotypic variance, and they were further assessed (Supplementary Table S11). There were 6 genes related to HSW within these QTL intervals, and 5 were annotated, including the CCCH-type zinc finger protein and calcium-binding protein. There were 53 seed shape-related genes and 30 genes were annotated, including serine/threonine phosphatase, bHLH transcription factor, calcium-binding protein Ca^{2+}/H^{+}-exchanging protein, and other functional genes. Seed color-related genes included 39 and 19 genes that were annotated, including ubiquitin-like protein, the WD40 family, and transcription factors. There were 79 genes associated with nutritional quality traits, and 41 genes were annotated, including numerous genes encoding enzymes, functional genes, and some transcription factors.

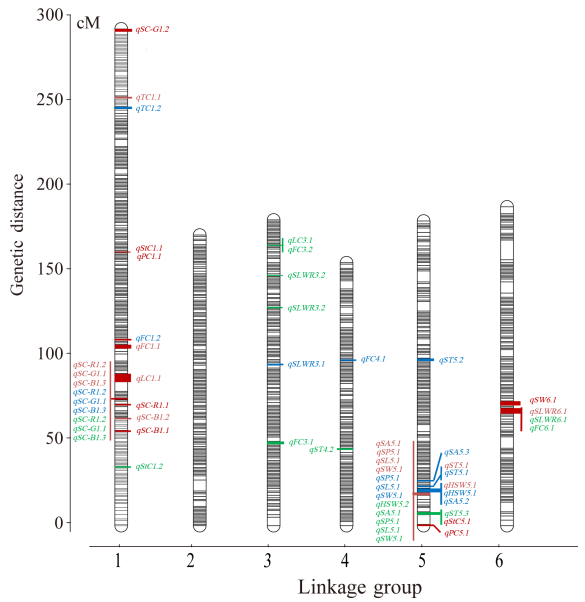


FIGURE 3
The ultra–high density genetic linkage map of faba bean based on bin makers and QTLs of seed-related traits. Note: Only the QTLs with the phenotypic variation > 10% were shown. Red words means QTLs detected in the F_2, blue words means QTLs detected in the $F_{2:3}$ of Jiuhua and green words means QTLs detected in the $F_{2:3}$ of Xueyao.

4.6 Reference genome mapping

Sequences of gene in our genetic map were well alignment with the recent published reference genome of faba bean (Supplementary

10.3389/fpls.2023.1201103

TABLE 2 Summary of the consensus reference genetic map of faba bean in this study.

Linkage groups	SNP count	Bin count	Length (cM)	Average interval (cM)	Largest gap size (cM)	Numbers of gaps > 5 cM
LG1	3,325	285	296.818	0.089	5.811	1
LG2	2,207	173	173.630	0.079	6.651	1
LG3	2,121	184	182.697	0.086	4.975	0
LG4	1,758	157	157.084	0.089	4.975	0
LG5	1,597	156	181.963	0.114	6.650	3
LG6	1,015	151	190.461	0.188	11.784	4
Total	12,023	1106	1182.653	0.098	11.784	9

TABLE 3 QTLs distribution of 15 seed-related traits with responsible for more than 10% of the explained phenotypic variation.

Trait	LG	QTL	environment	Left maker		Right maker		VG/VP (%)	Peak LOD	Add
				name	pos	name	pos			
HSW	5	*qHSW5.1*	F_2(2020)	yDN135233_c2_g1_2027	19.5	yDN120969_c0_g1_320	22.5	35.84	17.81	345.255
			$F_{2:3}$-JH(2021)	yDN135233_c2_g1_2027	19.5	yDN120969_c0_g1_320	22.5	38.88	15.79	214.079
		qHSW5.2	$F_{2:3}$-XY(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	25.71	10.42	228.937
SA	5	*qSA5.1*	F_2(2020)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	51.51	18.41	81.633
			$F_{2:3}$-XY(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	40.33	15.13	54.410
		qSA5.2	$F_{2:3}$-JH(2021)	yDN135233_c2_g1_2027	18.5	yDN120969_c0_g1_320	22.5	39.64	24.81	49.392
		qSA5.3	$F_{2:3}$-JH(2021)	hDN150254_c0_g5_96	26.5	dDN52935_c3_g4_215	28.5	15.19	11.34	5.348
SP	5	*qSP5.1*	F_2(2020)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	46.06	15.55	8.962
			$F_{2:3}$-JH(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	55.47	23.55	5.636
			$F_{2:3}$-XY(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	40.13	14.82	5.828
SL	5	*qSL5.1*	F_2(2020)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	42.65	14.06	2.724
			$F_{2:3}$-JH(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	48.32	28.30	1.756
			$F_{2:3}$-XY(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	35.96	13.79	1.904
SW	5	*qSW5.1*	F_2(2020)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	39.37	14.92	1.857
			$F_{2:3}$-JH(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	42.93	22.55	1.165
			$F_{2:3}$-XY(2021)	yDN151173_c1_g2_816	18.5	yDN135233_c2_g1_2027	19.5	36.78	12.90	1.356
	6	*qSW 6.1*	F_2(2020)	yDN128644_c0_g1_417	71.5	hDN154119_c0_g2_776	75.5	10.69	5.07	0.866
SLWR	3	*qSLWR 3.1*	$F_{2:3}$-JH(2021)	yDN154982_c0_g1_462	95.5	yDN150491_c1_g1_2729	96.5	18.23	6.37	0.004
		qSLWR 3.2	$F_{2:3}$-XY(2021)	hDN149791_c1_g1_342	129.5	yDN133005_c0_g1_512	130.5	17.24	11.11	-0.036
		qSLWR 3.3	$F_{2:3}$-XY(2021)	yDN155504_c1_g2_308	147.5	yDN155504_c1_g2_253	149.5	16.78	9.38	-0.036
	6	*qSLWR 6.1*	F_2(2020)	yDN145987_c0_g1_369	67.5	yDN138086_c0_g1_82	71.5	25.02	7.34	-0.046
			$F_{2:3}$-XY(2021)	yDN145987_c0_g1_369	67.5	yDN138086_c0_g1_82	70.5	10.40	6.21	0.032
ST	4	*qST4.1*	$F_{2:3}$-XY(2021)	hDN131761_c0_g1_1016	45.5	hDN122802_c0_g1_447	46.5	16.90	12.25	0.127

(Continued)

10.3389/fpls.2023.1201103

TABLE 3 Continued

Trait	LG	QTL	environment	Left maker		Right maker		VG/VP (%)	Peak LOD	Add
				name	pos	name	pos			
	5	qST5.1	F_2(2020)	yDN131562_c0_g2_1383	23.5	dDN54339_c2_g2_203	24.5	24.78	7.29	0.646
			$F_{2:3}$-JH(2021)	yDN131562_c0_g2_1383	23.5	dDN54339_c2_g2_203	24.5	19.19	9.38	0.452
		qST5.2	$F_{2:3}$-JH(2021)	hDN154491_c1_g8_222	98.5	hDN152331_c2_g3_311	99.5	10.07	4.78	0.123
		qST5.3	$F_{2:3}$-XY(2021)	hDN148575_c1_g1_723	7.5	yDN151173_c1_g2_935	9.5	10.66	8.58	0.475
SC-R	1	qSC-R1.1	F_2(2020)	hDN132853_c1_g2_224	71.5	dDN45140_c0_g1_2482	72.5	68.60	35.13	24.007
		qSC-R1.2	F_2(2020)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	12.90	11.02	11.943
			$F_{2:3}$-JH(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	73.99	34.76	35.411
			$F_{2:3}$-XY(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	65.35	29.45	34.151
SC-G	1	qSC-G1.1	F_2(2020)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	51.48	35.25	17.308
			$F_{2:3}$-JH(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	66.97	28.88	25.116
			$F_{2:3}$-XY(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	68.34	33.01	36.541
		qSC-G1.2	F_2(2020)	yDN157063_c3_g3_806	295.5	yDN157063_c3_g3_836	296	12.36	12.19	-11.891
SC-B	1	qSC-B1.1	F_2(2020)	yDN147029_c0_g1_601	56.5	yDN142452_c3_g4_331	57.5	24.62	15.73	9.483
		qSC-B1.2	F_2(2020)	hDN135643_c3_g1_514	63.5	yDN151467_c2_g1_440	64.5	22.55	16.11	7.153
		qSC-B1.3	F_2(2020)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	10.63	8.93	-6.273
			$F_{2:3}$-JH(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	37.04	12.09	9.231
			$F_{2:3}$-XY(2021)	hDN125239_c1_g4_1476	75.5	yDN127251_c0_g1_756	76.5	54.25	20.22	16.479
FC	1	qFC1.1	F_2(2020)	yDN145946_c2_g3_400	107.5	yDN125063_c0_g1_87	108.5	21.37	9.44	2.049
		qFC1.2	$F_{2:3}$-JH(2021)	dDN53089_c3_g1_46	109.5	yDN129665_c0_g3_238	110.5	14.85	6.26	1.379
	3	qFC3.1	$F_{2:3}$-XY(2021)	yDN148417_c0_g1_302	48.5	hDN142257_c0_g1_1594	49.5	15.43	5.97	1.057
		qFC3.2	$F_{2:3}$-XY(2021)	yDN119514_c0_g1_312	166.5	hDN145176_c0_g1_497	167.5	13.65	5.32	-0.962
	4	qFC4.1	$F_{2:3}$-JH(2021)	hDN122239_c0_g2_1049	98.5	hDN154311_c1_g1_631	99.5	10.16	4.97	1.008
	6	qFC6.1	$F_{2:3}$-XY(2021)	yDN145987_c0_g1_369	67.5	yDN138086_c0_g1_82	71.5	10.82	4.65	0.686
TC	1	qTC1.1	F_2(2020)	hDN124375_c0_g1_451	255.5	dDN47789_c0_g1_281	256.5	22.70	6.46	-0.061
		qTC1.2	$F_{2:3}$-JH(2021)	hDN122621_c4_g3_51	249.5	yDN154539_c0_g2_558	251.5	17.61	5.17	-0.034
StC	1	qStC1.1	F_2(2020)	dDN40232_c0_g1_396	162.5	yDN134012_c0_g1_800	163.5	13.77	8.19	-1.306
		qStC1.2	$F_{2:3}$-XY(2021)	hDN148143_c3_g2_213	34.5	yDN141447_c5_g1_240	35.5	18.44	5.53	1.041
	5	qStC5.1	F_2(2020)	hDN148575_c1_g1_723	7.5	yDN151173_c1_g2_935	9.5	11.38	6.64	-1.192
PC	1	qPC1.1	F_2(2020)	dDN40232_c0_g1_396	162.5	yDN134012_c0_g1_800	163.5	17.32	6.12	1.070
	5	qPC5.1	F_2(2020)	dDN41265_c0_g1_1106	0	dDN41265_c0_g1_1104	0.5	13.65	4.99	0.875
LC	1	qLC1.1	F_2(2020)	hDN155223_c0_g1_2012	86.5	hDN146106_c2_g1_2134	88.5	20.78	5.91	-0.141
	3	qLC3.1	$F_{2:3}$-XY(2021)	yDN119514_c0_g1_312	166.5	hDN145176_c0_g1_497	167.5	21.35	12.72	-0.130

The QTL with underlines means stable QTL for each trait. HSW 100-seed weight (g), SA seed surface area (mm^2), SP seed perimeter (mm), SL seed length (mm), SW seed width (mm), SLWR seed length and width ratio, ST seed thickness (mm), SC-R seed coat color R value, SC-G seed coat color G value, SC-B seed coat color B value, FC fiber content (%), TC tannin content (%), StC starch content (%), PC protein content (%), LC lipid content (%), JH Jiuhua, XY Xueyao.

Table S12). It was found that about 60% of the genes in each LG were mapped to the corresponding chromosome. Specifically, LG1–LG6 were assigned to chromosome 1, chromosome 3, chromosome 2, chromosome 5, chromosome 4 and chromosome 6, respectively.

Candidate genes were mapped to the reference genome and most annotated genes were located on other five chromosomes except the chromosome 3 (Supplementary Figure S1; Supplementary Table 11). Twenty-five genes were located on chromosome 1L (the long arm of

TABLE 4 Details of genes and SNPs of 15 seed-related traits in QTL interval based on 130K TNGS.

Trait	Total QTL number	SNP number	Gene number	VG/VP >10 QTLs interval Gene number	Stable QTLs interval gene number
HSW	3	15	8	6	5
SA	5	23	13	9	2
SP	1	4	2	2	2
SL	4	37	20	2	2
SW	5	31	22	9	2
SLWR	5	38	20	15	2
ST	8	61	40	23	10
SC-R	3	74	30	25	13
SC-G	5	88	42	14	13
SC-B	4	63	28	21	13
FC	7	28	20	17	3
TC	2	91	49	49	0
StC	7	30	21	10	0
PC	2	5	4	4	0
LC	4	22	15	7	0
Total	65	610	333	213	67

HSW 100-seed weight (g), SA seed surface area (mm^2), SP seed perimeter (mm), SL seed length (mm), SW seed width (mm), SLWR seed length and width ratio, ST seed thickness (mm), SC-R seed coat color R value, SC-G seed coat color G value, SC-B seed coat color B value, FC fiber content (%), TC tannin content (%), StC starch content (%), PC protein content (%), LC lipid content (%).

chromosome 1) and 17 genes were located on chromosome 1S (the short arm of chromosome 1). Seven, eleven, one and seven of these annotated genes were located on chromosome 2, chromosome 4, chromosome 5, chromosome 6, respectively. Furthermore, there were also seven genes located on free chromosomes (the unassigned scaffolds that cannot be placed on any known chromosome).

5 Discussion

5.1 The first ultra-dense genetic map for faba bean

Owing to the rapid development of high-throughput sequencing technologies, sufficient molecular markers can now be obtained to facilitate the mapping of high-density genetic maps and research on map-based gene cloning (Yang et al., 2012; Zhang et al., 2016; Zhou et al., 2018; Gaur et al., 2020; Gu et al., 2020; Sa et al., 2021). The molecular genetic analysis of faba bean is currently lagging in comparison to that of many other crops due to its large genome size (Adhikari et al., 2021). Establishing a reliable linkage map between genetic markers and traits is one of the key approaches to improve molecular breeding without a reference genome (Chapman et al., 2022). In this study, a genetic map of faba beans was constructed using high-throughput genotyping platforms. To date, genetic map construction using microarray chips has been successfully reported in several crops, such as pea (Tayeh et al., 2015), wheat (Liu et al., 2018; Ren et al., 2021), cotton (Gu et al., 2020) and pepper (Cheng et al., 2016). In addition, the 130 K liquid-phase gene chip used in this study was developed using transcriptome data, which contains large-scale information. Furthermore, all marker sequences provided valuable gene information, indicating that this liquid-phase gene chip is an effective and feasible tool to utilize for genetic map construction.

There have been more than 20 genetic maps reported for faba beans. Of these, the genetic map constructed by Carrillo-perdomo et al. (2020) containing 1,728 markers, with a total length of 1,547.71 cM and an average genetic distance of 0.89 cM. To date, one of the two SNP genetic map constructed by Li et al. (2023) had the highest density, containing 5,103 markers, with a total length of 1,333.31 cM and an average genetic distance of 0.26 cM. In the present study, an ultra-dense genetic map was constructed, encompassing 12,023 markers in 6 LGs, with an average distance of 0.098 cM. The number, density, and distribution quality of the new molecular markers was thus significantly higher when compared with previous genetic maps. The presented genetic map only has 9 gaps > 5 cM, and thus, it can be effectively utilized for faba bean gene mapping and MAS breeding.

5.2 Comparison with previous QTL reports

QTL mapping and the analysis of candidate genes within QTL intervals is an effective strategy to investigate numerous crop traits

(Bornowski et al., 2020; Chen et al., 2021), and can contribute to the development of molecular marker-assisted breeding (Torres et al., 2010). The 100-seed weight, seed shape, and nutritional quality of faba beans are all quantitative traits susceptible to environmental influence. To improve the accuracy of QTL mapping for seed traits, QTL analysis was performed in the F_2 and $F_{2:3}$ populations in two locations. There was a total of 65 seed trait-related QTLs detected (Supplementary Table S6), of which, 11 were repeatedly detected in different environments (Supplementary Table S9).

Patto et al. (1999) used a genetic map constructed using the F_2 population and found that most of the QTLs related to seed weight were located on chromosome 6 for faba bean. Using the recombinant in bred line (RIL6) population constructed using Vf6 and Vf27, Ávila et al. (2017) identified 5 QTLs for HSW, which were located on 4 different chromosomes. Tian et al. (2018) identified two QTLs for seed weight using an F_2 population derived from Yun122/TF42, which were located on two different LGs. In this study, we identified three QTLs linked to HSW, one at LG4, and two at LG5. These results indicate that faba bean seed weight is controlled by multiple main-effect QTLs. *qHSW5.1*, one of the three QTLs related to HSW, was also associated with SA, and *qHSW5.2* was associated with SA, SP, SL, and SW, which indicated that these two QTLs are also involved in controlling seed shape (Table 3).

Seed shape traits are among the most important factors used to determine seed size. The localization and cloning of seed shape genes are of great importance when aiming to increase crop yield and improve appearance quality (Austin and Lee, 1996; Song et al., 2007; Verma et al., 2015; Cheng et al., 2017; Murube et al., 2020). According to the Gramene website (http://archive.gramene.org/qtl/), more than 400 rice grain shape-related genes/QTLs have been identified through genetic mapping and correlation analysis. However, few studies have reported QTL mapping for the seed shape traits of faba bean, a seed length-related QTL and a seed width-related QTL were identified by Tian et al. (2018), 8 QTLs related to seed length, 9 QTLs related to seed width and 8 QTLs related to seed thickness were identified by Li et al. (2023). In this investigation, 28 QTLs for 6 seed-shape traits were identified using linkage analysis, and most were located on LG5(Table 3; Supplementary Table S6). Compared to these QTLs reported, those identified as controlling seed shape in this study were new, and could thus be applied to the subsequent fine mapping of seed shape traits and the investigation of related genes in faba bean. *qSA5.1*, *qSLWR6.1* and *qST5.1* were stable QTLs explained > 10% of phenotypic variation, while *qSA5.1* was also associated with SP, SL, and SW. which indicated that these QTLs can be used for further fine mapping and superior gene discovery of seed shape traits.

Seed coat color is a key factor affecting seed quality (Yoshimura et al., 2012; García-Fernández et al., 2021). Different seed coat colors may have different functions (Debeaujon et al., 2003), and the different seed coat colors of faba beans may also be associated with different nutritional qualities. The results of the correlation analysis among seed traits showed that seed coat color was positively correlated with FC and StC, and negatively correlated with LC. Mendel first proposed that the seed coat color of peas is controlled by a pair of genes and considered a qualitative trait (Myers, 2004), while the seed coat color of soybeans is controlled by multiple genetic loci (Choung et al., 2001), and more than 30 molecular marker loci on different chromosomes that control seed coat color in soybean have been detected (Yuan et al., 2022). However, few studies on the QTLs for seed coat color in faba bean have been reported. WY7 and TCX7, the parents used in this study, have purple and white coats, respectively. A total of 12 QTLs, mainly located on LG1, were detected by quantitative measurement of the SC-R, SC-G, and SC-B. *qSC-R1.2* was also located with SC-G and SC-B (Table 3), which could explain the > 50% phenotypic variation. *qSC-R1.1* is located with SC-G, and *qSC-R1.3* is located with SC-B (Supplementary Table S6). These three QTLs are key objects for further study of grain coat color traits.

The main nutrients in faba bean seeds are protein and starch, with low lipid and fiber content levels, as well as tannin (Zanotto et al., 2020), pyrimidine glucoside, and other bioactive substances (Björnsdotter et al., 2021). QTL mapping for quality traits can help to improve the utilization and value of faba beans. At present, there are relatively few studies on the QTL mapping of quality traits in broad beans. Only five genes that control grain proteins have been identified (Macas et al. 1993b). In this study, 22 QTLs linked to quality traits were detected using SNP markers for the first time, including 7 QTLs for FC, 7 for StC, 4 for LC, 2 for PC, and 2 for TC (Supplementary Table S6). In particular, *qFC1.1*, *qTC1.1*, *qLC1.1*, and *qLC3.1* could explain > 20% of the phenotypic variation, and *qStC1.1* was also associated with PC (Table 3). These QTLs could thus be used to identify the candidate genes for faba bean quality traits.

5.3 Candidate genes for the QTLs controlling seed-related traits

To identify candidate genes for seed-related traits in faba bean, we focused on 213 genes within 41 QTL intervals that explained > 10% of the phenotypic variation. According to the results of the functional annotation, 57.28% of these genes had been annotated. Signaling pathways that regulate seed size in plants include the ubiquitin-protease pathway, mitogen-activated protein kinase signaling pathway, transcriptional regulation, G-protein signaling pathway, IKU pathway, and plant hormones (Gnan et al., 2014; Li and Li, 2016; Li et al., 2019). Jayakodi et al. (2023) identified 15 marker–seed size associations, and most prominent signal was located on chromosome 4 within the *Vfaba.Hedin2.R1.4g051440* gene. In this investigation, there were 30 genes annotations among the 59 genes linked to HSW and seed shape (Supplementary Table S11). Thirteen of these annotation genes located on chromosome 4 by whole genome sequence alignment. dou_TRINITY_DN52935_c3_g4 and hua_TRINITY_DN119282_c0_g1 encode serine/threonine phosphatase and the transcription factor bHLH, respectively, which are reportedly involved in regulating seed size (Savadi, 2018). dou_TRINITY_DN38848_c0_g1 encodes a *CYP* gene and *CYP* is involved in protein folding, signal transduction, and RNA processing (Krücken et al., 2009). There are also two calcium signaling pathway genes, including a calcium-binding protein gene ye_TRINITY_DN120969_c0_g1 and a Ca^{2+}/H^{+}-

exchanging protein gene hua_TRINITY_DN154119_c0_g2, which may be involved in the Ca signaling pathway to regulate seed development. Other unannotated candidate genes could also potentially regulate seed size.

The seed coat color of plants is affected by numerous factors, but flavonoids are the decisive pigments (Lepiniec et al., 2006). In this study, there were 34 candidate genes for seed coat color, 19 of which were annotated (Supplementary Table S11). Among these genes, the translated product of ye_TRINITY_DN150431_c0_g1 is a ubiquitin-like protein that plays an important role in pigment accumulation (Tang et al., 2015). ye_TRINITY_DN150347_c0_g1 and ye_TRINITY_DN139828_c0_g1 are WD40 family genes, which have been suggested to regulate the formation of proanthocyanidins in seed coats (Shirley et al., 1995; Walker et al., 1999). Furthermore, the other 16 annotated genes and 15 unannotated genes may also be required for the pigment composition of different seed coat colors, but this requires further verification.

In this study, 79 candidate genes were associated with five nutritional quality traits, of which, 41 were annotated (Supplementary Table S11). There were 7 genes for LC, 4 of which were annotated, but no functions related to lipid synthesis and accumulation were reported. hua_TRINITY_DN145176_c0_g1, a crude fiber candidate gene, is a triose-phosphate transporter gene that reportedly affects starch and glucose transport in transgenic tobacco (Häusler et al., 1998). dou_TRINITY_DN53089_c3_g1 and ye_TRINITY_DN155843_c1_g1 are GDSL esterases that may also be involved in fiber metabolism. Condensed tannins, also known as proanthocyanidins, exhibit antioxidant, antibacterial, anticancer, and anti-mutation activities (Gutierrez et al., 2020). The two genes zt-1 and zt-2 are the most studied for controlling tannin content in faba bean (Gutierrez et al., 2006; Gutierrez et al., 2007; Gutierrez et al., 2008). Of the candidate genes related to tannins, dou_TRINITY_DN58315_c1_g1 encodes a bHLH transcription factor gene, which is reportedly involved in the mechanisms of tannin biosynthesis in faba bean (Gutierrez et al., 2020). Other tannin-annotated genes obtained in the target intervals have not been reported in faba bean, and thus may be candidate genes affecting tannin content. Further studies are required to confirm the functions of these genes.

5.4 Reference genome mapping analysis

Compared to chromosomes and gene locations of the reference genome, the number of linkage groups in our genetic map was consistent with their respective chromosomes, but there were variations in the order of genes on the chromosome, and about 25% of them were not found in the genome (Supplementary Table S12). Eighty-five candidate genes within the QTL interval were mapped to the reference genome, seven of which were located on the contigs (Supplementary Figure S1). Therefore, a part of contigs on the reference genome can be assembled to the genome of faba bean based on the map constructed in this study, which is conducive to the further improvement of the physical map of faba bean.

6 Conclusions

A high-density genetic map with 12,023 SNPs in 6 LGs was constructed using the faba_bean_ 130 K SNP TNGS genotyping platform. A total of 65 QTLs for seed-related traits were identified (3 for 100-seed weight, 28 for seed shape, 12 for seed coat color, and 22 for nutritional quality). Furthermore, 333 candidate genes were identified that are likely to participate in the regulation of seed-related traits. This is the first ultra-dense genetic map of faba bean and it provides a foundation for further genetic analyses, MAS breeding, and reference genome assembly research. This study will also be useful for faba bean gene isolation and functional genomics research.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

NZ and LW conceived and designed the experiments. DX, YM, YW, YZ, MY and CG performed experiment, EZ, KW and BL analysed data. NZ wrote the manuscript. LW and XW revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This study was financially supported by Jiangsu Agricultural Science and Technology Innovation Fund (CX(22)3144), The Fund of China Agriculture Research System (CARS-08-Z10), Jiangsu Province Seed Industry Revitalization Project (JBGS(2021)056), Jiangsu 333 Talent Project (SRCB202210-01) and Research Fund for the Doctoral Program of Jiangsu Yanjiang Institute of Agricultural Sciences (YJBS(2021)001).

Acknowledgments

We would like to thank Editage (www.editage.cn) for english language editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1201103/full#supplementary-material

References

Adhikari, K. N., Khazaei, H., Ghaouti, L., Maalouf, F., Vandenberg, A., Link, W., et al. (2021). Conventional and molecular breeding tools for accelerating genetic gain in faba bean (*Vicia faba* l.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.744259

Aguilar–Benitez, D., Casimiro–Soriguer, I., Maalouf, F., and Torres, A. M. (2021). Linkage mapping and QTL analysis of flowering time in faba bean. *Sci. Rep.* 11, 13716. doi: 10.1038/s41598-021-92680-4

Alghamdi, S. S., Migdadi, H. M., Ammar, M. H., Paull, J. G., and Siddique, K. H. M. (2012). Faba bean genomics: current status and future prospects. *Euphytica* 186, 609–624. doi: 10.1007/s10681-012-0658-4

Amarowicz, R., and Shahidi, F. (2017). Antioxidant activity of faba bean seed extract and its phenolic composition. *J. Func Foods* 38, 656–662. doi: 10.1016/j.jff.2017.04.002

Arbaoui, M., Link, W., Satovic, Z., and Torres, A. M. (2008). Quantitative trait loci of frost tolerance and physiologically related trait in faba bean (*Vicia faba* l.). *Euphytica* 164, 93–104. doi: 10.1007/s10681-008-9654-0

Austin, D. F., and Lee, M. (1996). Comparative mapping in $F_{2:3}$ and $F_{6:7}$ generations of quantitative trait loci for grain yield and yield components in maize. *Theor. Appl. Genet.* 92, 817–826. doi: 10.1007/BF00221893

Ávila, C. M., Ruiz-Rodríguez, M. D., Cruz-Izquierdo, S., Atienza, S. G., Cubero, J. I., and Torres, A. M. (2017). Identification of plant architecture and yield-related QTL in *Vicia faba* l. *Mol. Breed.* 37, 88. doi: 10.1007/s11032-017-0688-7

Ávila, C. M., Satovic, Z., Sillero, J. C., Nadal, S., Rubiales, D., Moreno, M. T., et al. (2005). QTL detection for agronomic traits in faba bean (*Vicia faba* l.). *Agric. Conspec Sci.* 70, 65–73.

Ávila, C. M., Satovic, Z., Sillero, J. C., Rubiales, D., Moreno, M. T., and Torres, A. M. (2004). Isolate and organ-specific QTLs for ascochyta blight resistance in faba bean (*Vicia faba* l). *Theor. Appl. Genet.* 108, 1071–1078. doi: 10.1007/s00122-003-1514-7

Björnsdotter, E., Nadzieja, M., Chang, W., Escobar-Herrera, L., Mancinotti, D., Angra, D., et al. (2021). VC1 catalyses a key step in the biosynthesis of vicine in faba bean. *Nat. Plants* 7, 923–931. doi: 10.1038/s41477-021-00950-w

Bornowski, N., Song, Q. J., and Kelly, J. D. (2020). QTL mapping of post-processing color retention in two black bean populations. *Theor. Appl. Genet.* 133, 3085–3100. doi: 10.1007/s00122-020-03656-3

Carrillo-Perdomo, E., Magnin-Robert, J. B., Raffiot, B., Deulvot, C., Floriot, M., Lejeune-Hénaut, I., et al. (2022). A QTL approach in faba bean highlights the conservation of genetic control of frost tolerance among legume species. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.970865

Carrillo-perdomo, E., Vidal, A., Kreplak, J., Duborjal, H., Leveugle, M., Duarte, J., et al. (2020). Development of new genetic resources for faba bean (*Vicia faba* l.) breeding through the discovery of gene-based SNP markers and the construction of a high-density consensus map. *Sci. Rep.* 10, 6790. doi: 10.1038/s41598-020-63664-7

Catt, S. C., Braich, S., Kaur, S., and Paull, J. G. (2017). QTL detection for flowering time in faba bean and the responses to ambient temperature and photoperiod. *Euphytica* 213, 125. doi: 10.1007/s10681-017-1910-8

Chapman, M. A., He, Y. Q., and Zhou, M. L. (2022). Beyond a reference genome: pangenomes and population genomics of underutilized and orphan crops for future food and nutrition security. *New Phytol.* 234, 1583–1597. doi: 10.1111/nph.18021

Chen, H., Pan, X. W., Wang, F. F., Liu, C. K., Wang, X., Li, Y. S., et al. (2021). Novel QTL and meta-QTL mapping for major quality traits in soybean. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.774270

Cheng, R. R., Kong, Z. G., Zhang, L. W., Xie, Q., Jia, H. Y., Yu, D., et al. (2017). Mapping QTLs controlling kernel dimensions in a wheat inter-varietal RIL mapping population. *Theor. Appl. Genet.* 130, 1405–1414. doi: 10.1007/s00122-017-2896-2

Cheng, J. W., Qin, C., Tang, X., Zhou, H. K., Hu, Y. F., Zhao, Z. C., et al. (2016). Development of a SNP array and its application to genetic mapping and diversity assessment in pepper (*Capsicum* spp.). *Sci. Rep.* 6, 33293. doi: 10.1038/srep33293

Choung, M. G., Baek, I. Y., Kang, S. T., Han, W. Y., Shin, D. C., Moon, H. P., et al. (2001). Isolation and determination of anthocyanins in seed coats of black soybean (*Glycine max* (L.) merr.). *J. Agric. Food Chem.* 49, 5848–5851. doi: 10.1021/jf010550w

Cruz-Izquierdo, S., Ávila, C. M., Satovic, Z., Palomino, C., Gutierrez, N., Ellwood, S. R., et al. (2012). Comparative genomics to bridge vicia faba with model and closely-related legume species: stability of QTLs for flowering and yield-related traits. *Theor. Appl. Genet.* 125, 1767–1782. doi: 10.1007/s00122-012-1952-1

Cubero, J. I. (1974). On the evolution of *Vicia faba* l. *Theor. Appl. Genet.* 45, 47–51. doi: 10.1007/BF00283475

Debeaujon, I., Nesi, N., Perez, P., Devic, M., Grandjean, O., Caboche, M., et al. (2003). Proanthocyanidin-accumulating cells in arabidopsis testa: regulation of differentiation and role in seed development. *Plant Cell* 15, 2514–2531. doi: 10.1105/tpc.014043

Díaz-Ruiz, R., Satovic, Z., Ávila, C. M., Alfaro, C. M., Gutierrez, M. V., Torres, A. M., et al. (2009). Confirmation of QTLs controlling ascochyta fabae resistance in different generations of faba bean (*Vicia faba* l.). *Crop Pasture Sci.* 60, 353–361. doi: 10.1071/CP08190

Díaz-Ruiz, R., Torres, A. M., Satovic, Z., Gutiérrez, M. V., Cubero, J. I., and Román, B. (2010). Validation of QTLs for *Orobanche crenata* resistance in faba bean (*Vicia faba* l.) across environments and generations. *Theor. Appl. Genet.* 120, 909–919. doi: 10.1007/s00122-009-1220-1

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Ellwood, S. R., Phan, H. T. T., Jordan, M., Hane, J., Torres, A. M., Avila, C. M., et al. (2008). Construction of a comparative genetic map in faba bean (*Vicia faba* l.); conservation of genome structure with *Lens culinaris*. *BMC Genomics* 9, 380. doi: 10.1186/1471-2164-9-380

Etemadia, F., Hashemi, M., Barker, A. V., Zandvakili, R. O., and Liu, X. B. (2019). Agronomy,nutritional value, and medicinal application of faba bean (*Vicia faba* l.). *Hortic. Plant J.* 5, 170–182. doi: 10.1016/j.hpj.2019.04.004

García-Fernández, C., Campa, A., and Ferreira, J. J. (2021). Dissecting the genetic control of seed coat color in a RIL population of common bean (*Phaseolus vulgaris* l.). *Theor. Appl. Genet.* 134, 3687–3698. doi: 10.1007/s00122-021-03922-y

Gaur, R., Verma, S., Pradhan, S., Ambreen, H., and Bhatia, S. (2020). A high-density SNP-based linkage map using genotyping-by-sequencing and its utilization for improved genome assembly of chickpea (*Cicer arietinum* l.). *Funct. Integr. Genomics* 20, 763–773. doi: 10.1007/s10142-020-00751-y

Gnan, S., Priest, A., and Kover, P. X. (2014). The genetic basis of natural variation in seed size and seed number and their trade-off using *Arabidopsis thaliana* MAGIC lines. *Genetics* 198, 1751–1758. doi: 10.1534/genetics.114.170746

Gu, Q. S., Ke, H. F., Liu, Z. W., Lv, X., Sun, Z. W., Zhang, M., et al. (2020). A high –density genetic map and multiple environmental tests reveal novel quantitative trait loci and candidate genes for fibre quality and yield in cotton. *Theor. Appl. Genet.* 133, 3395–3408. doi: 10.1007/s00122-020-03676-z

Gutierrez, N., Avila, C. M., Duc, G., Marget, P., Suso, M. J., Moreno, M. T., et al. (2006). CAPs markers to assist selection for low vicine and convicine contents in faba bean (*Vicia faba* l.). *Theor. Appl. Genet.* 114, 59–66. doi: 10.1007/s00122-006-0410-3

Gutierrez, N., Avila, C. M., Moreno, M. T., and Torres, A. M. (2008). Development of SCAR markers linked to zt-2, one of genes controlling absence of tannins in faba bean. *Aust. J. Agr Res.* 59, 62–68. doi: 10.1071/AR07019

Gutierrez, N., Avila, C. M., Rodriguez-Suarez, C., Moreno, M. T., and Torres, A. M. (2007). Development of SCAR markers linked to a gene controlling absence of tannins in faba bean. *Mol. Breed.* 19, 305–314. doi: 10.1007/s11032-006-9063-9

Gutierrez, N., Avila, C. M., and Torres, A. M. (2020). The bHLH transcription factor VfTT8 underlies zt2, the locus determining zero tannin content in faba bean (*Vicia faba* l.). *Sci. Rep.* 10, 14299. doi: 10.1038/s41598-020-71070-2

Gutiérrez, N., Palomino, C., Satovic, Z., Ruiz-Rodríguez, M. D., Vitale, S., Gutiérrez, M. V., et al. (2013). QTLs for orobanche spp. resistance in faba bean: identification and validation across different environments. *Mol. Breed.* 32, 909–922. doi: 10.1007/s11032-013-9920-2

Gutierrez, N., and Torres, A. M. (2021). QTL dissection and mining of candidate genes for ascochyta fabae and orobanche crenata resistance in faba bean (*Vicia faba* l.). *BMC Plant Biol.* 21, 551. doi: 10.1186/s12870-021-03335-5

Häusler, R. E., Schlieben, N. H., Schulz., B., and Flügge, U.-I. (1998). Compensation of decreased triose phosphate/phosphate translocator activity by accelerated starch turnover and glucose transport in transgenic tobacco. *Planta* 204, 366–376. doi: 10.1007/s004250050268

Hou, W. W., Zhang, X. J., Yan, Q. B., Li, P., Sha, W. C., Tian, Y. Y., et al. (2018). Linkage map of a gene controlling zero tannins (*zt-1*) in faba bean (*Vicia faba* l.) with SSR and ISSR markers. *Agronmy* 8, 80. doi: 10.3390/agronomy8060080

Jayakodi, M., Golicz, A. A., Kreplak, J., Fechete, L. I., Angra, D., Bednář, P., et al. (2023). The giant diploid faba genome unlocks variation in a global protein crop. *Nature* 615 (7953), 652–659. doi: 10.1038/s41586-023-05791-5

Jensen, E. S., Peoples, M. B., and Hauggaard-Nielsen, H. (2010). Faba bean in cropping systems. *Field Crop Res.* 115, 203–216. doi: 10.1016/j.fcr.2009.10.008

Johnston, J. S., Bennett, M. D., Rayburn, A. L., Galbraith, D. W., and Price, H. J. (1999). Reference standards for determination of DNA content of plant nuclei. *Am. J. Bot.* 86, 609–613. doi: 10.2307/2656569

Kaur, S., Kimber, R. B. E., Cogan, N. O. I., Materne, M., Forster, J. W., and Paull, J. G. (2014). SNP discovery and high-density genetic mapping in faba bean (*Vicia faba* l.) permits identification of QTLs for ascochyta blight resistance. *Plant Sci.*, 217, 47–55. doi: 10.1016/j.plantsci.2013.11.014

Khazaei, H., O'Sullivan, D. M., Sillanpää, M. J., and Stoddard, F. L. (2014). Use of synteny to identify candidate genes underlying QTL controlling stomatal traits in faba bean (*Vicia faba* l.). *Theor. Appl. Genet.* 127, 2371–2385. doi: 10.1007/s00122-014-2383-y

Krücken, J., Greif, G., and Samson-Himmelstjerna, G. (2009). In silico analysis of the cyclophilin repertoire of apicomplexan parasites. *Parasit Vectors* 2, 27. doi: 10.1186/1756-3305-2-27

Lepiniec, L., Debeaujon, I., Routaboul, J. M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006). Genetics and biochemistry of seed flavonoids. *Annu. Rev. Plant Biol.* 57, 405–430. doi: 10.1146/annurev.arplant.57.032905.105252

Li, C., Duan, Y. H., Miao, H. M., Ju, M., Wei, L. B., and Zhang, H. Y. (2021). Identification of candidate genes regulating the seed coat color trait in sesame (*sesamum indicum* l.) using an integrated approach of QTL mapping and transcriptome analysis. *Front. Genet.* 12. doi: 10.3389/fgene.2021.700469

Li, M. W., He, Y. H., Liu, R., Li, G., Wang, D., Ji, Y. S., et al. (2023). Construction of SNP genetic map based on targeted next-generation sequencing and QTL mapping of vital agronomic traits in faba bean (*Vicia faba* l.). *J. Integr. Agric.* 22, 2095-3119 doi: 10.1016/j.jia.2023.01.003

Li, N., and Li, Y. H. (2016). Signaling pathways of seed size control in plants. *Curr. Opin. Plant Biol.* 33, 23–32. doi: 10.1016/j.pbi.2016.05.008

Li, N., Xu, R., and Li, Y. H. (2019). Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* 70, 435–463. doi: 10.1146/annurev-arplant-050718-095851

Liu, J. J., Luo, W., Qin, N. N., Ding, P. Y., Zhang, H., Yang, C. C., et al. (2018). A 55 K SNP array–based genetic map and its utilization in QTL mapping for productive tiller number in common wheat. *Theor. Appl. Genet.* 131, 2439–2450. doi: 10.1007/s00122-018-3164-9

Ma, Y., Bao, S. Y., Yang, T., Hu, J. G., Guan, J. P., He, Y. H., et al. (2013). Genetic linkage map of Chinese native variety faba bean (*Vicia faba* l.) based on simple sequence repeat markers. *Plant Breed.* 132, 397–400. doi: 10.1111/pbr.12074

Macas, J., Dolezel, J., Lucretti, S., Pich, U., Meister, A., Fuchs, J., et al. (1993a). Localization of seed protein genes on flow-sorted field bean chromosomes. *Chromosome Res.* 1, 107–115. doi: 10.1007/BF00710033

Macas, J., Weschke, W., Bümlein, H., Pich, U., Houben, A., Wobus, U., et al. (1993b). Localization of vicilin genes *via* polymerase chain reaction on microisolated field bean chromosomes. *Plant J.* 3, 883–886. doi: 10.1111/j.1365-313X.1993.00883.x

Martineau-Côté, D., Achouri, A., Karboune, S., and L'Hocine, L. (2022). Faba bean: an untapped source of quality plant proteins and bioactives. *Nutrients* 14, 1541. doi: 10.3390/nu14081541

Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in bi-parental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Murube, E., Campa, A., Song, Q. J., McClean, P., and Ferreira, J. J. (2020). Toward validation of QTLs associated with pod and seed size in common bean using two nested recombinant inbred line populations. *Mol. Breed.* 40, 7. doi: 10.1007/s11032-020-01155-3

Myers, J. R. (2004). An alternative possibility for seed coat color determinaton in mendel's experiment. *Genetics* 166, 1137. doi: 10.1534/genetics.166.3.1137

Ocaña–Moral, S., Gutiérrez, N., Torres, A. M., and Madrid, E. (2017). Saturation mapping of regions determining resistance to ascochyta blight and broomrape in faba bean using transcriptome–based SNP genotyping. *Theor. Appl. Genet.* 130, 2271–2282. doi: 10.1007/s00122-017-2958-5

Patto, M. C. V., Torres, A. M., Koblizkova, A., Macas, J., and Cubero, J. I. (1999). Development of a genetic composite map of *Vicia faba* using F_2 populations derived from trisomic plants. *Theor. Appl. Genet.* 98, 736–743. doi: 10.1007/s001220051129

Rastas, P. (2017). Lep-MAP3: robust linkage mapping even for low-coverage whole genome sequencing data. *Bioinformatics* 33 (23), 3726–3732. doi: 10.1093/bioinformatics/btx494

Ren, T. H., Fan, T., Chen, S. L., Li, C. S., Chen, Y. Y., Ou, X., et al. (2021). Utilization of a Wheat55K SNP array–derived high–density genetic map for high–resolution mapping of quantitative trait loci for important kernel–related traits in common wheat. *Theor. Appl. Genet.* 134, 807–821. doi: 10.1007/s00122-021-03765-7

Román, B., Satovic, Z., Pozarkova, D., Macas, J., Dolezel, J., Cubero, J. I., et al. (2004). Development of a composite map in vicia faba, breeding applications and future prospects. *Theor. Appl. Genet.* 108, 1079–1088. doi: 10.1007/s00122-003-1515-6

Román, B., Torres, A. M., Rubiales, D., Cubero, J. I., and Satovic, Z. (2002). Mapping of quantitative trait loci controlling broomrape (*Orobanche crenata forsk.*) resistance in faba bean (*Vicia faba* l.). *Genome* 45, 1057–1063. doi: 10.1139/g02-082

Sa, K. J., Choi., I. Y., Park, J. Y., Choi, J. K., Ryu, S. H., and Lee, J. K. (2021). Mapping of QTL for agronomic traits using high–density SNPs with an RIL population in maize. *Genes Genom* 43, 1403–1411. doi: 10.1007/s13258-021-01169-x

Sallam, A., Arbaoui, M., El-Esawi, M., Abshire, N., and Martsch, R. (2016). Identification and verification of QTL associated with frost tolerance using linkage mapping and GWAS in winter faba bean. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01098

Satovic, Z., Ávila, C. M., Cruz-Izquierdo, S., Díaz-Ruíz, R., García-Ruíz, G. M., Palomino, C., et al. (2013). A reference consensus genetic map for molecular markers and economically important traits in faba bean (*Vicia faba* l.). *BMC Genomics* 14, 932. doi: 10.1186/1471-2164-14-932

Satovic, Z., Torres, A. M., and Cubero, J. I. (1996). Genetic mapping of new morphological, isozyme and RAPD markers in *Vicia faba* l. using trisomics. *Theor. Appl. Genet.* 93, 1130–1138. doi: 10.1007/BF00230136

Savadi, S. (2018). Molecular regulation of seed development and strategies for engineering seed size in crop plants. *Plant Growth Regul.* 84, 401–422. doi: 10.1007/s10725-017-0355-3

Shirley, B. W., Kubasek, W. L., Storz, G., Bruggemann, E., Koornneef, M., Ausubel, F. M., et al. (1995). Analysis of arabidopsis mutants deficient in flavonoid biosynthesis. *Plant J.* 8, 659–671. doi: 10.1046/j.1365-313x.1995.08050659.x

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Sudheesh, S., Kimber, R. B. E., Braich, S., Forster, J. W., Paull, J. G., and Kaur, S. (2019). Construction of an integrated genetic linkage map and detection of quantitative trait loci for ascochyta blight resistance in faba bean (*Vicia faba* l.). *Euphytica* 215, 42. doi: 10.1007/s10681-019-2365-x

Tang, X. F., Miao, M., Niu, X. L., Zhang, D. F., Cao, X. L., Jin., X. C., et al. (2015). Ubiquitin-conjugated degradation of golden 2-like transcription factor is mediated by CUL4-DDB1-based E3 ligase complex in tomato. *New Phytol.* 209, 1028–1039. doi: 10.1111/nph.13635

Tayeh, N., Aluome, C., Falque, M., Jacquin, F., Klein, A., Chauveau, A., et al. (2015). Development of two major resources for pea genomics: the GenoPea 13.2K SNP array and a high-density, high-resolution consensus genetic map. *Plant J.* 84, 1257–1273. doi: 10.1111/tpj.13070

Tian, Y. Y., Hou, W. W., and Liu, Y. J. (2018). Genetic analysis and QTL mapping for seed traits in broad bean. *Mol. Plant Breed.* 16 (4), 1174–1183. doi: 10.13271/j.mpb.016.001174

Torres, A. M., Avila, C. M., Gutierrez, N., Palomino, C., Moreno, M. T., and Cubero, J. I. (2010). Marker-assisted selection in faba bean (*Vicia faba* l.). *Field Crops Res.* 115, 243–252. doi: 10.1016/j.fcr.2008.12.002

Torres, A. M., Weeden, N. F., and Martín, A. (1993). Linkage among isozyme, RFLP and RAPD markers in vicia faba. *Theor. Appl. Genet.* 85, 937–945. doi: 10.1007/BF00215032

Verma, P., Goyal, R., Chahota, R. K., Sharma, T. R., Abdin, M. Z., and Bhatia, S. (2015). Construction of a genetic linkage map and identification of QTLs for seed weight and seed size traits in lentil (*Lens culinaris medik.*). *PloS One* 10, e0139666. doi: 10.1371/journal.pone.0139666

Walker, A. R., Davison, P. A., Bolognesi-Winfield, A. C., James, C. M., Srinivasan, N., Blundell, T. L., et al. (1999). The TRANSPARENT TESTA GLABRA1 locus, which regulates trichome differentiation and anthocyanin biosynthesis in arabidopsis, encodes a WD40 repeat protein. *Plant Cell* 11, 1337–1350. doi: 10.1105/tpc.11.7.1337

Wang, S., Basten, C. J., and Zeng, Z. B. (2012). *Windows QTL cartographer 2.5* (Raleigh, NC: Department of Statistics, North Carolina State University). Available at: http://statgen.ncsu.edu/qtlcart/WQTLCart.htm.

Wang, C. Y., Liu, R., Liu, Y. J., Hou, W. W., Wang, X. J., Miao, Y. M., et al. (2021). Development and application of the Faba_bean_130K targeted next-generation sequencing SNP genotyping platform based on transcriptome sequencing. *Theor. Appl. Genet.* 134, 3195–3207. doi: 10.1007/s00122-021-03885-0

Wang, H. F., Zong, X. X., Guan, J. P., Yang, T., Sun, X. L., Ma, Y., et al. (2012). Genetic diversity and relationship of global faba bean (*Vicia faba* l.) germplasm revealed by ISSR markers. *Theor. Appl. Genet.* 1249, 789–797. doi: 10.1007/s00122-011-1750-1

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., et al. (2016). A SNP-based consensus genetic map for synteny-based trait targeting in faba bean (*Vicia faba* l.). *Plant Biotechnol. J.* 14, 177–185. doi: 10.1111/pbi.12371

Yang, T., Bao, S. Y., Ford, R., Jia, T. J., Guan, J. P., He, Y. H., et al. (2012). High-throughput novel microsatellite marker of faba bean *via* next generation sequencing. *BMC Genomics* 13, 602. doi: 10.1186/1471-2164-13-602

Yang, T., Jiang, J. Y., Zhang, H. Y., Liu, R., Strelkov, S., Hwang, S. F., et al. (2019). Density enhancement of a faba bean genetic linkage map (*Vicia faba*) based on simple sequence repeats markers. *Plant Breed.* 138, 207–215. doi: 10.1111/pbr.12679

Yoshimura, Y., Zaima, N., Moriyama, T., and Kawamura, Y. (2012). Different localization patterns of anthocyanin species in the pericarp of black rice revealed by imaging mass spectrometry. *PloS One* 7, e31285. doi: 10.1371/journal.pone.0031285

Yuan, B. Q., Yuan, C. P., Wang, Y. M., Liu, X. D., Qi, G. X., Wang, Y. N., et al. (2022). Identification of genetic loci conferring seed coat color based on a high-density map in soybean. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.968618

Zanotto, S., Khazaei, H., Elessawy, F. M., Vandenberg, A., and Purves, R. W. (2020). Do faba bean genotypes carrying different zero-tannin genes (zt1 and zt2) differ in phenolic profiles? *J. Agric. Food Chem.* 68, 7530–7540. doi: 10.1021/acs.jafc.9b07866

Zhang, H. Y., Miao, H. M., Li, C., Wei, L. B., Duan, Y. H., Ma, Q., et al. (2016). Ultra-dense SNP genetic map construction and identification of SiDt gene controlling the determinate growth habit in *Sesamum indicum* l. *Sci. Rep.* 6, 31556. doi: 10.1038/srep31556

Zhou, G. F., Jian, J. B., Wang, P. H., Li, C. D., Tao, Y., Li, X., et al. (2018). Construction of an ultra-high density consensus genetic map, and enhancement of the physical map from genome sequencing in *Lupinus angustifolius*. *Theor. Appl. Genet.* 131, 209–223. doi: 10.1007/s00122-017-2997-y

Zong, X. X., Liu, X. J., Guan, J. P., Wang, S. M., Liu, Qc., Paull, J. G., et al. (2009). Molecular variation among Chinese and global winter faba bean germplasm. *Theor. Appl. Genet.* 118, 971–978. doi: 10.1007/s00122-008-0954-5

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 September 2023
DOI 10.3389/fpls.2023.1254365

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Zanping Han,
Henan University of Science and Technology, China
Zhang JiingBo,
Xinjiang Normal University, China
Hossein Sabouri,
Gonbad Kavous University, Iran

*CORRESPONDENCE
Juan Huang
huang200669@163.com
Shuxun Yu
yushuxun@zafu.edu.cn
Zhen Feng
fengzhen@zafu.edu.cn

†These authors have contributed equally to this work

RECEIVED 07 July 2023
ACCEPTED 18 August 2023
PUBLISHED 01 September 2023

CITATION
Li L, Hu Y, Wang Y, Zhao S, You Y, Liu R, Wang J, Yan M, Zhao F, Huang J, Yu S and Feng Z (2023) Identification of novel candidate loci and genes for seed vigor-related traits in upland cotton (*Gossypium hirsutum* L.) via GWAS. *Front. Plant Sci.* 14:1254365. doi: 10.3389/fpls.2023.1254365

Identification of novel candidate loci and genes for seed vigor-related traits in upland cotton (*Gossypium hirsutum* L.) via GWAS

Libei Li[1†], Yu Hu[1†], Yongbo Wang[2†], Shuqi Zhao[3], Yijin You[1], Ruijie Liu[1], Jiayi Wang[1], Mengyuan Yan[1], Fengli Zhao[4], Juan Huang[5*], Shuxun Yu[1*] and Zhen Feng[1*]

[1]The Key Laboratory for Quality Improvement of Agricultural Products of Zhejiang Province, College of Advanced Agricultural Sciences, Zhejiang A&F University, Lin'an, Hangzhou, China, [2]Cotton Sciences Research Institute of Hunan, Changde, Hunan, China, [3]Cotton and Wheat Research Institute, Huanggang Academy of Agricultural Sciences, Huanggang, Hubei, China, [4]State Key Laboratory of Rice Biology and Breeding, China National Rice Research Institute, Hangzhou, China, [5]Research Center of Buckwheat Industry Technology, Guizhou Normal University, Guiyang, China

Seed vigor (SV) is a crucial trait determining the quality of crop seeds. Currently, over 80% of China's cotton-planting area is in Xinjiang Province, where a fully mechanized planting model is adopted, accounting for more than 90% of the total fiber production. Therefore, identifying SV-related loci and genes is crucial for improving cotton yield in Xinjiang. In this study, three seed vigor-related traits, including germination potential, germination rate, and germination index, were investigated across three environments in a panel of 355 diverse accessions based on 2,261,854 high-quality single-nucleotide polymorphisms (SNPs). A total of 26 significant SNPs were detected and divided into six quantitative trait locus regions, including 121 predicted candidate genes. By combining gene expression, gene annotation, and haplotype analysis, two novel candidate genes (*Ghir_A09G002730* and *Ghir_D03G009280*) within *qGR-A09-1* and *qGI/GP/GR-D03-3* were associated with vigor-related traits, and *Ghir_A09G002730* was found to be involved in artificial selection during cotton breeding by population genetic analysis. Thus, understanding the genetic mechanisms underlying seed vigor-related traits in cotton could help increase the efficiency of direct seeding by molecular marker-assisted selection breeding.

KEYWORDS

upland cotton, seed vigor, germination rate, GWAS, candidate genes

Introduction

Upland cotton (*Gossypium hirsutum* L.) is one of the world's most important cash crops and a major source of natural fibers, accounting for more than 95% of global cotton production (Chen et al., 2007). Lint yield depends largely on the quality of cotton seeds, while seed vigor (SV) is crucial for evaluating seed quality (Sawan, 2016). SV also determines the growth of crops and food safety; for example, rapidly and uniformly germinating seeds can significantly increase the emergence rate in the field and suppress weed growth (He et al., 2019a). In addition, with the widespread application of mechanized direct seeding (DS) in cotton production, cotton seeds with low vigor will make it difficult to sow all seedlings at once, leading to many problems such as uneven seedling age and weak seedling vigor (Qun et al., 2007; Liu et al., 2015). Therefore, the identification of loci and genes related to SV is urgently needed for DS of cotton.

Seed germination is a key factor affecting SV traits in plants. Phytohormones such as gibberellin (GA) and abscisic acid (ABA) have been reported to be essential for the regulation of seed germination (Yamaguchi, 2008; Ryu and Cho, 2015)—for example, GA and ABA synthesis pathway-related genes (*GA20ox3, GA3ox1*, *GA2ox5*, *ABI3*, and *ABI5*) have a strong effect on seed germination (Yamauchi et al., 2004; Yamaguchi, 2008; Iglesias-Fernandez and Matilla, 2009). When plants are under abiotic stress, ABA in the plant will increase rapidly, and high levels of ABA will close the stomata and activate complex signaling pathways mediated by kinase/phosphatase regulation (Kim et al., 2010). Low levels of reactive oxygen species (ROS) act as signaling particles to promote dormancy release and trigger seed germination (Li et al., 2022)—for example, *OsCDP3.10* promotes the accumulation of H_2O_2 during the early stage of seed germination by increasing the amino acid content (Peng et al., 2022). The relationship between seed germination and the ROS scavenging system has been validated in many crops and other plants, such as *Arabidopsis* (Leymarie et al., 2012), wheat (Ishibashi et al., 2008), and rice (Ye et al., 2012). Furthermore, crosstalks between ABA and ROS signaling pathways have also been reported in plants. In rice, *qSE3* significantly increased ABA biosynthesis and activated ABA signaling responses, resulting in decreased H_2O_2 levels in germinating seeds under salinity stress (He et al., 2019b).

SV-related traits are quantitative traits controlled by both genetic and environmental factors (Li W. et al., 2021). These traits include germination rate (GR), germination percentage (GP), germination index (GI), vigor index (VI), seedling shoot length (SL), and shoot fresh weight (FW) (Dai et al., 2022; Si et al., 2022). In recent years, linkage mapping has been widely used to identify SV-related quantitative trait loci (QTLs) in crops, and multiple QTLs have been cloned (Fujino et al., 2004; Fujino et al., 2008; He et al., 2019b; Jiang et al., 2020; Veisi et al., 2022). By using BC_1F_5 populations derived from a rice intraspecific cross ('WTR-1' × 'Y134'), 28 SV-related QTLs were identified by a SNP genotyping array, and one major QTL ($q1stGC_{11.2}$) explaining 19.9% of the phenotypic variation (PV) was flanked by SNP_11_27994133 on chromosome 11 (Dimaano et al., 2020). In wheat, a total of 49 QTLs were detected on 12 chromosomes, including seven SV candidate genes involved in the processes of cell division during germination of aged seeds, carbohydrate and lipid metabolism, and transcription (Shi et al., 2020). Wang L. et al. (2022) constructed a linkage map based on specific-locus-amplified fragment sequencing (SLAF-seq) SNP markers in melon; *2020/2021-qsg5.1* was significant in both environments, and *MELO3C031219.2*, in this region, exhibited a significant expression difference between the parental lines during multiple germination stages (Wang L. et al., 2022). Under low temperature conditions, three QTLs (*qLTG-3-1*, *qLTG3-2*, and *qLTG-4*) related to GR were identified by 122 backcross inbred lines, and the phenotypic variation explained (PVE) by *qLTG-3-1* was 35.0% (Fujino et al., 2004). Subsequently, *qLTG-3-1* was cloned, which was closely related to tissue vacuolation, by covering the embryo (Fujino et al., 2008). Furthermore, the genome-wide association study (GWAS) approach is a method in which germplasm resources are used to study the genetic structure of target traits. Compared to traditional QTL mapping, GWAS can provide higher resolution by using ancestral recombination events and has been successfully applied to identify significant SNP loci and potential candidate genes associated with important agronomic traits in major crops (Zhu et al., 2008; Shikha et al., 2021)—for example, SV-related QTLs were identified in 346 rice accessions using GWAS, while 51 significant SNPs were detected for SL, GR, and FW (Dai et al., 2022). In addition, a previous study involving 187 rice accessions identified the candidate gene *OsSAP16*; the loss of *OsSAP16* function reduced the rice seed germination rate (Wang et al., 2018). Recently, a candidate gene (*Gh_A09G1509*) responsible for seed germination was detected through a GWAS panel in upland cotton by using whole-genome resequencing (Si et al., 2022). These results suggest that genome-wide association analysis is an effective method for identifying genes associated with seed germination.

To date, many quantitative traits have been reported in cotton, such as fiber quality traits (Su et al., 2016b; Zhang et al., 2019), early maturity traits (Li et al., 2017; Li L. et al., 2021), and yield component traits (Su et al., 2016a; Feng et al., 2022). However, SV-related traits in cotton have received little attention, and most research have focused on seed germination in relation to stress tolerance (Yuan et al., 2019; Chen L. et al., 2020; Gu et al., 2021; Guo et al., 2022). Few candidate genes for cotton SV-related traits have been identified (Si et al., 2022), and the mechanism of seed germination needs further study. In this study, GR, GP, and GI were determined in a natural population of upland cotton in three environments, and whole-genome resequencing was used to achieve deep coverage and obtain high-quality SNP markers. In addition, six stable QTLs and two novel candidate genes (*Ghir_A09G002730* and *Ghir_D03G009280*) for SV-related traits were further identified by a GWAS panel, laying the foundation for understanding the genetic mechanism underlying SV and providing potential information for applying these potential elite loci for marker-assisted selection (MAS) in cotton breeding.

Materials and methods

GWAS population and field experiments

The 355 upland cotton germplasm resources collected by laboratories worldwide represent a natural population. Previous studies focused on early maturity (Li L. et al., 2021), fiber quality (Su et al., 2016b), fiber yield (Su et al., 2016a; Feng et al., 2022), and plant architecture component traits based on abundant phenotypic variations in this population (Su et al., 2018). These upland cotton varieties are from different countries and represent accessions resulting from more than 100 years of global upland cotton breeding. Seeds of the GWAS population used for phenotyping SV-related traits were collected from three environments, including Huanggang in Hubei Province (30°57′ N, 114°92′ E) in 2021 (E1: Huanggang-2021) and Sanya in Hainan Province (18°36′ N, 109° 17′ E) in two consecutive years (2021 and 2022) (E2: Sanya-2021 and E3: Sanya-2022). The field experiments in Sanya and Huanggang were conducted following a randomized complete block design with two and three replications, respectively.

Phenotyping for SV-related traits and statistical analysis

The phenotyping of SV-related traits was carried out by the sandponic method based on previously described methods (Si et al., 2022). Cotton seeds collected from the field were ginned, and cotton fuzz was removed by concentrated sulfuric acid. Then, all seeds were sun-dried for 2 days to break dormancy uniformly. A total of 150 plump seeds with uniform size and full grain were selected, disinfected with 15% sodium hypochlorite for 10 min, and then washed clean with distilled water. Then, each line was evenly planted in a plastic sand box containing 800 g of dry quartz sand with a size of 13 cm × 19 cm × 12 cm. Subsequently, the seeds were covered with 250 g of dry quartz sand, and 200 mL of distilled water was added. The number of germinated seeds was counted each day until the seventh day. All experiments were conducted in a phytotron with 16 h of light (25°C) and 8 h of darkness (18°C). Three biological replicates were included for each accession, and 50 seeds were used for each replicate. Moreover, three SV-related traits (GR, GP, and GI) were selected for measurement. The full name, abbreviation, and measurement method of each trait are listed in Table 1 as described by Yuan et al. (2019). The statistical analysis of the maximum value, minimum value, average value, etc., was performed using R software (version: 4.2.2).

Development of SNP markers

The resequencing data (PRJNA389777) of the 355 upland cotton germplasms used in this study were reported in a previous study (Li L. et al., 2021). The Illumina HiSeq4000 platform was used for paired-end read sequencing, with an average sequencing depth of more than 10×. Based on previously released data, the new variation map of the natural population was employed in the 'HaplotypeCaller' module of GATK (version: 4.2.6.1) (Mckenna et al., 2010). Briefly, the variation detection process was as follows: (1) The quality of paired-end reads from 355 accessions was assessed using FastQC (version: 0.11.9) (Andrews, 2010); (2) Sequencing quality control was carried out with fastp software (version: 0.23.2) to obtain high-quality reads with the following parameters: '-w 16 -c -l 80 -5 -3 -W 4 -M 20 -f 10 -F 13 -t 3 -T 3 -q 20 -u 40' (Chen et al., 2018); (3) All high-quality reads were mapped to the 'TM-1' (version: HAU_v1.1) reference genome using BWA (version: 0.7.17-r1188) (Li, 2013; Wang et al., 2019); (4) Then, Picard software (https://github.com/broadinstitute/picard) was used to sort the BAM file and mark duplicate reads; (5) The 'HaplotypeCaller' module of GATK (version: 4.2.6.1) was used to identify variant sites and perform SNP filtering with the following conditions: 'QUAL <30, DP <1,340, DP >10,050, QD <2.0, MQ <35, FS >70, SOR >3, MQRankSum <-12.5, and ReadPosRankSum <-4.0'; (6) The SNP clusters with at least three SNPs detected within a 10-base window were removed; (7) SNPs within five base pairs of an InDel were filtered out by BCFtools software (version: 0.1.19-44428cd) (Danecek et al., 2021); and (8) SNPs with a minor allele frequency (MAF) <5% and missing rate <20% were discarded by VCFtools (version: 0.1.16) (Danecek et al., 2011).

GWAS and genetic diversity analysis

Genome-wide association analysis was performed by combining 2,262,367 high-quality SNPs with the phenotype data of 355 upland cotton accessions collected in three environments for SV-related traits using linear mixed models in GEMMA (version: 0.98.3) and executed by vcf2gwas software (version: 0.8.7) (Zhou and Stephens, 2012; Vogt et al., 2022). $P < 1 \times 10^{-6}$ was used as the threshold to detect significant SNP loci. Additionally, the PVE by

TABLE 1 Method of measurement for seed vigor-related traits.

Trait	Trait abbreviation	Measurement methods for each trait
Germination potential	GP	The number of germinated seeds in the early stage of germination (3 days)/the number of seeds tested
Germination rate	GR	The number of germinated seeds on the 7th day after planting/the number of tested seeds
Germination index	GI	GI = Σ(Gt/Dt), where Gt represents the number of germinated seeds per day and Dt represents the number of days corresponding to Gt

each marker was calculated as previously reported (Feng et al., 2022). The nucleotide diversity (π)a value was calculated using VCFtools based on the release years (before the 1950s, 1950s–1970s, 1980s–1990s, and 2000s–2020s) and geographical distribution (early maturity region: NSER, Yellow River region: YRR, Yangtze River region: YZRR, and Northwest Inland region: NIR) of the 355 accessions. The packages 'CMplot' (https://github.com/YinLiLin/CMplot), 'LDheatmap' (Shin et al., 2006), and 'ggplot2' (Wickham, 2011) in R software were used to generate Manhattan plots and for linkage disequilibrium (LD) block analysis and haplotype analysis.

Candidate gene identification and expression analysis

Based on the 'TM-1' reference genome (HAU_v1.1) (Wang et al., 2019), the genes in the interval located 200 kb upstream and downstream of the significant SNPs were defined as candidate genes. The protein sequences of the candidate genes were obtained from Cottongene (https://www.cottongen.org/). Then, local BLAST software was used to compare the protein sequence of the candidate gene with the *Arabidopsis* protein database (https://www.arabidopsis.org) to obtain the homologous sequence, and the criterion was set to less than E^{-60} (Johnson et al., 2008). The expression patterns of SV candidate genes in upland cotton were determined by RNA-seq and quantitative reverse-transcription PCR (qRT-PCR) analysis. RNA isolation method was performed as described by Feng et al. (2022). *GhUBQ7* was used as an internal control. Quantitative analysis method was performed using a Roche real-time qPCR system (Light Cycler 480 II) and SYBR with three biological repeats. The public RNA-seq data (PRJNA248163) including SRR1695160, SRR1695161, and SRR1695162 were downloaded from NCBI (https://www.ncbi.nlm.nih.gov/bioproject/). The Illumina Hiseq2000 platform was used to perform RNA sequencing on 'TM1' seeds soaked in water for 0, 5, and 10 h, and the paired-end clean reads length was more than 100 bp. The gene expression values were normalized by the average expression levels (log2) based on transcripts per million values. The clustered heat map was drawn by the R package 'pheatmap' (Kolde, 2012).

Results

Characterization and distribution of SNPs in the upland cotton genome

Resequencing of the natural population libraries by the Illumina HiSeq 4000 platform with 150 bp paired-end reads, as described in previous reports (Li L. et al., 2021), yielded approximately 65,013 million reads in total for the 355 cotton genotypes. Approximately 88.3% of the total bases were successfully mapped to the cotton reference genome, and the statistical sequencing depth corresponded to 11.7-fold in the 355 upland cotton accessions. A total of 2,262,367 SNPs distributed across the cotton genome with a MAF >0.05, and missing rate of resequencing data of less than 20% was used for the GWAS of the 355 cotton germplasm accessions, of which the At and Dt subgenomes contained 1,404,637 and 857,730 SNPs, resulting in an average SNP density of 993.44 and 1045.91 SNP/Mb, respectively (Table 2; Figure 1). The percentage of the SNPs in each chromosome varied from 1.4% on chromosome D04 to 11.4% on chromosome A08 (Figure 1). Most of the SNPs were

TABLE 2 Distribution and frequency of single-nucleotide polymorphisms (SNPs) identified using the resequencing approach in upland cotton.

Chromosome	Chromosome length (Mb)	SNP number	Density (SNP/Mb)	Chromosome	Chromosome length (Mb)	SNP number	Density (SNP/Mb)
A01	117.76	102,597	871.25	D01	63.21	97,337	1,539.92
A02	108.09	56,850	525.94	D02	69.84	86,010	1,231.56
A03	113.06	73,858	653.27	D03	52.70	37,138	704.70
A04	85.15	48,890	574.16	D04	56.43	33,068	586.00
A05	109.42	93,469	854.23	D05	62.93	49,985	794.25
A06	124.06	216,693	1,746.73	D06	66.87	95,435	1,427.18
A07	97.78	82,817	846.95	D07	59.26	85,111	1,436.29
A08	122.38	259,187	2,117.94	D08	69.04	93,091	1,348.38
A09	82.10	82,034	999.16	D09	52.82	74,192	1,404.64
A10	114.85	102,498	892.44	D10	68.01	59,948	881.51
A11	123.21	85,696	695.52	D11	72.94	44,642	612.02
A12	107.67	65,645	609.67	D12	62.69	55,606	886.94
A13	108.38	134,403	1,240.15	D13	63.34	46,167	728.84
Total	1,413.91	1,404,637	993.44	Total	820.08	857,730	1,045.91

10.3389/fpls.2023.1254365

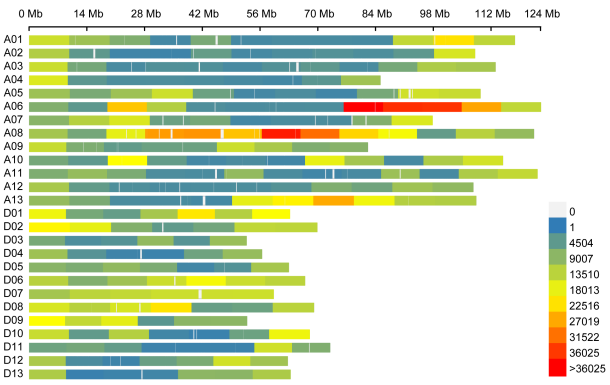


FIGURE 1
Single-nucleotide polymorphism distributions in the upland cotton genome. The number of SNPs within a 10-Mb window. A01–A13 and D01–D13 on the Y axis are the numbers of the 26 chromosomes. The X axis represents chromosome length (Mb).

located in intergenic regions (84.38%), whereas the exonic and intronic genome regions contained only 0.89% and 3.03% of SNPs, respectively (Supplementary Table S1). In addition, SNPs in the coding regions (coding sequences, CDSs) included 33.26% synonymous mutations and 64.13% nonsynonymous mutations.

PV of SV-related traits

The three SV-related traits (GI, GP, and GR) of natural population accessions were measured in three environments. The values followed a normal distribution for the GI and GP but showed a skewed distribution for GR based on Shapiro–Wilk tests (Supplementary Table S2). The frequency histograms of SV-related traits are shown in Figures 2A–I. The lowest average GI was 55.23 in the E1 environment, and the highest average GI was 58.92 in the E2 environment, with a coefficient of variation (CV) ranging from 6.52% to 12.02% (Supplementary Table S2). For GP, the E1 environment had the lowest average value of 70.92%, while the E2 environment had the highest average value of 79.61%; the CV in the E1 environment (11.67%) was higher than that in the E2 environment (9.21%) and the E3 environment (9.37%) (Supplementary Table S2). For GR, the lowest average value was 87.37% in the E1 environment, and the highest average value was 93.25% in the E2 environment, with a CV ranging from 3.71% to 10.48% (Supplementary Table S2). Two-way analysis of variance (ANOVA) showed that genotype (G) and the genotype-by-environment interaction (G × E) had significant effects on the GI, GP, and GR ($P < 0.001$) (Supplementary Table S3). Furthermore, the heritability of these three SV-related traits ranged from 74.23% (GR) to 81.75% (GP), whereas that of GI was 76.03% (Supplementary Table S3). These results suggested that SV-related traits have extensive PV in the GWAS panel, which is suitable for further GWAS.

GWAS of SV-related Traits in Upland Cotton

A total of 292 significant SNPs for three SV-related traits were identified on 11 chromosomes using the linear mixed model (Figure 3; Supplementary Table S4; Supplementary Figures S1-S3). Only 11 SNPs were identified in the At subgenome, whereas 281 SNPs were localized to the Dt subgenome. Among them, chromosome D03 had the highest number of SNPs (281), with a total of 254, and the range of $-\log_{10}(p)$ values was from 6.00 to 8.27. Furthermore, 26 stable SNPs were identified in a minimum of two environments (including for the best linear unbiased predictor,

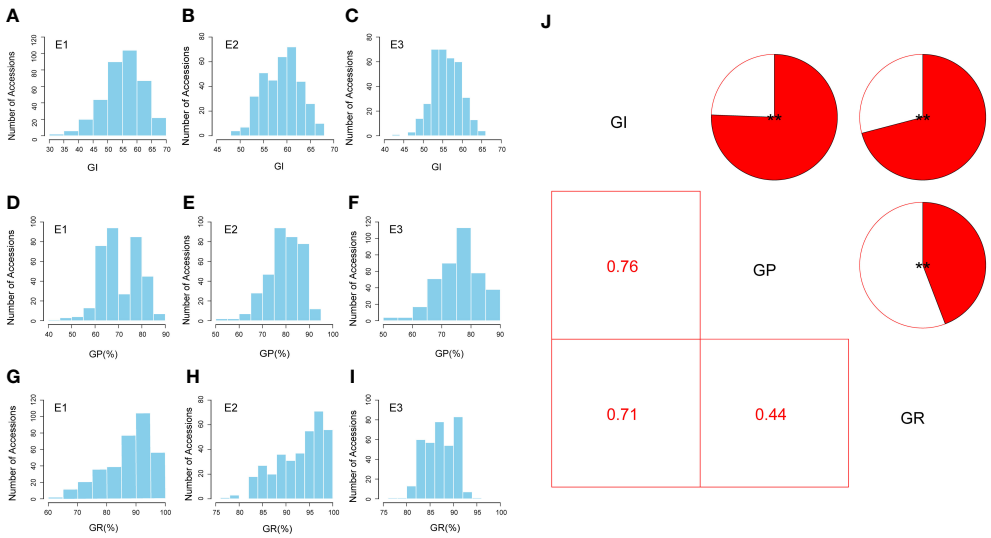


FIGURE 2
Phenotypic variation analysis of seed vigor (SV)-related traits. **(A–I)** Distributions of the mean values for the germination index (GI), germination potential (GP), and germination rate (GR) in three environments, respectively. **(J)** Correlation analysis of SV-related traits (GI, GP, and GR) in three environments (**$P < 0.01$). E1, Huanggang-2021; E2, Sanya-2021; and E3, Sanya-2022.

10.3389/fpls.2023.1254365

BLUP) or two traits, which were declared as six stable QTLs, focusing on chromosomes A09, A10, and D03. Notably, a QTL region (*qGR-A09-1*) located on chromosome A09 showed a strong SNP cluster associated with GR, which had a PVE of 6.76–8.56% and -$\log_{10}(P)$ ranging from 6.19 to 7.74. *qGP-A10-1* on chromosome A10 had only one SNP that explained 8.15% of the observed PVE, with a LOD score of 7.39. Four QTLs on chromosome D03 (*qGR/GI-D03-1*, *qGI/GR-D03-2*, *qGI/GP/GR-D03-3*, and *qGI/GP/GR-D03-4*) were identified in two, three, three, and four environments, explaining 6.61–7.39%, 6.72–7.79%, 6.65–8.43%, and 6.61–8.90% of the observed PVE, respectively. Interestingly, a stable QTL (*qGI/GP/GR-D03-3*) region was revealed on chromosome D03 from 31.68 to 32.61 Mb and was flanked by regions associated with the GI, GP, and GR in the E1, E3, and BLUP environments. Thus, the QTLs *qGR-A09-1* and *qGI/GP/GR-D03-3* could be treated as major QTLs for further dissection.

Identification of a candidate gene for GR on chromosome A09

In this study, a novel QTL, *qGR-A09-1*, exhibited a significant SNP cluster (rsA09_7745467, rsA09_7791621, rsA09_7878527, rsA09_7908017, rsA09_7954329, rsA09_7954353, and rsA09_7962794) occupying a physical region of 0.2 Mb on chromosome A09 (Figure 4A). Meanwhile, 22 genes were annotated in this QTL region based on the *G. hirsutum* reference genome (Wang et al., 2019), except for *Ghir_A09G002720* and *Ghir_A09G002760*, which did not have annotation information (Supplementary Table S5). We further conducted LD analysis on the significant SNP rsA09_7962794, and LD blocks were found in this region (Figure 4A). In this QTL interval, rsA09_7962794 on chromosome A09 showed a strong association with GR, with 7.95% of the PVE downstream of *Ghir_A09G002730* (Table 3). rsA09_7962794 had two haplotypes, GG and AA, which resulted in the accessions carrying the AA genotype having a significantly higher GR than those carrying the GG haplotype in three environments ($P < 0.01$) (Figure 4B). In addition, to gain a further understanding of the genetic characteristics of rsA09_7962794 in relation to geographic distribution, the 355 upland cotton accessions were divided into four groups: NIR, YZRR, YRR, and NSER. Interestingly, YRR and NSER showed an extraordinarily low frequency of the nonfavorable haplotype (GG), while the accessions obtained from YZRR and NIR had a relatively high frequency of the favorable haplotype (AA) (>75%) (Figure 4C). Furthermore, the genetic diversity of *Ghir_A09G002730* decreased following the breeding period. Cotton accessions released before the 1980s showed greater diversity than accessions bred from the 1980s

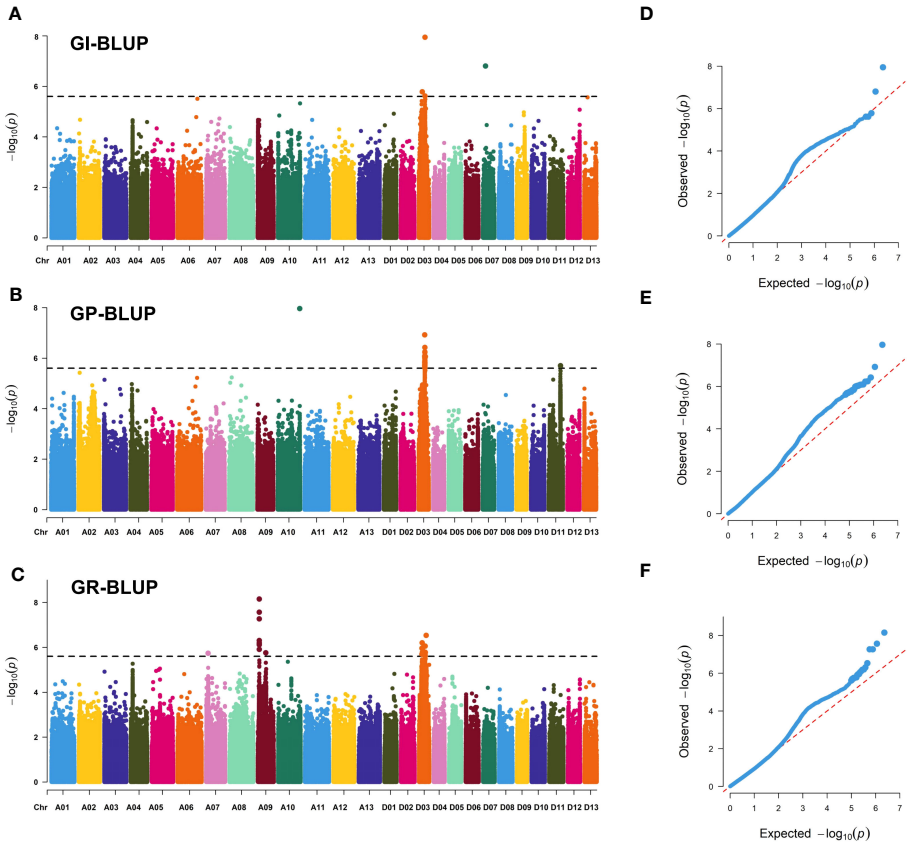


FIGURE 3
Genome-wide association study results for seed vigor-related traits. **(A–C)** Manhattan plots of GI-BLUP, GP-BLUP, and GR-BLUP for single-nucleotide polymorphism (SNP) markers, respectively. Significant SNP markers are distinguished by black lines. **(D–F)** QQ plots for GI-BLUP, GP-BLUP, and GR-BLUP, respectively.

10.3389/fpls.2023.1254365

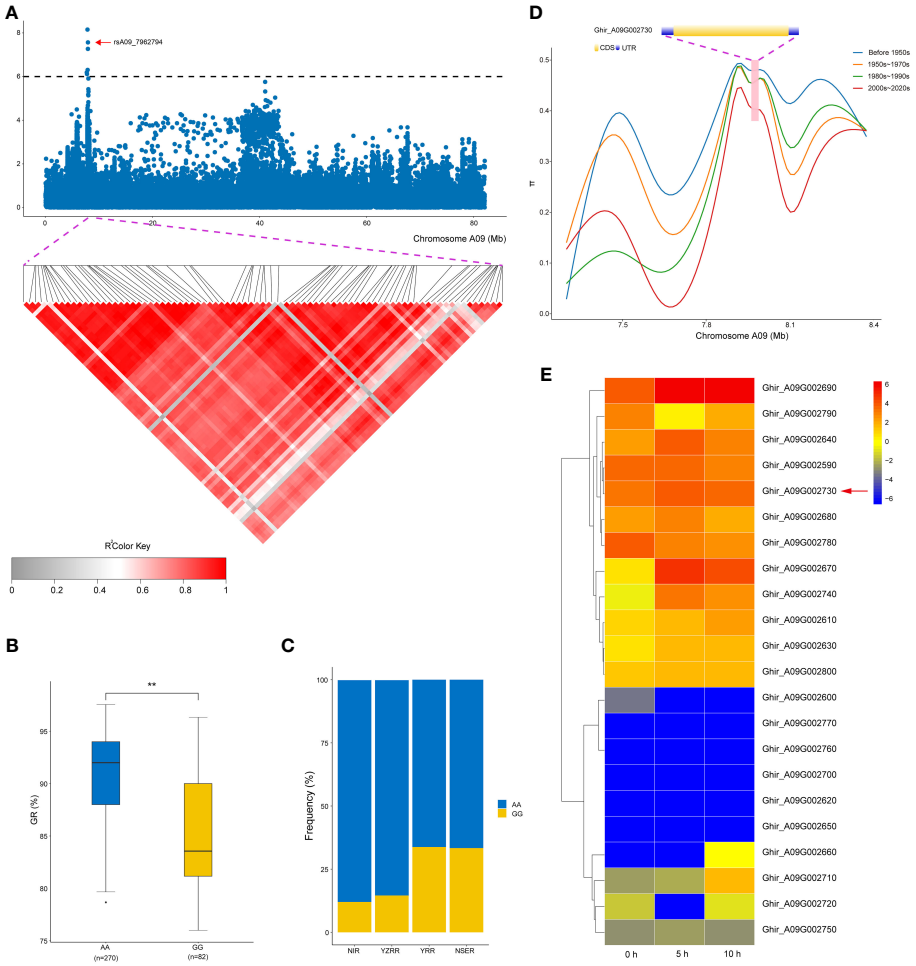


FIGURE 4
Variation analysis of the germination rate (GR)-associated gene Ghir_A09G002730 in the candidate region. **(A)** Local Manhattan plots for GR-related genes on chromosome A09 and linkage disequilibrium heat map for the candidate region within 21.9 kb. **(B)** Box plots for GR of the two haplotypes mentioned above (**$P < 0.01$). **(C)** Differentiation of the genetic diversity distribution of *Ghir_A09G002730* in four geographic areas (NIR, Northwest Inland region; YZRR, Yangtze River region; YRR, Yellow River region; and NSER, Northern Specific Early-Maturity region). **(D)** Gene structure diversity of *Ghir_A09G002730* across three breeding stages. **(E)** Heat map of candidate gene expression patterns in the seed germination stage (0, 5, and 10 h) on chromosome A09.

to the 2000s, while accessions bred after the 2000s showed the lowest diversity (Figure 4D). Specifically, *Ghir_A09G002730* belongs to the pentatricopeptide repeat (PPR) superfamily protein family and has higher expression levels during the seed germination stage from 0 to 10 h than other genes (Figure 4E). The qRT-PCR analysis also showed that *Ghir_A09G002730* had higher expression levels in the accessions ('Liaomian27' and 'Xinluzhong35') carrying the AA allele than in accessions ('PB12-1-8' and 'Xiazao2') with GG allele during the seed germination stage (Supplementary Figure S4).

Identification of a candidate gene for GR on chromosome D03

As mentioned above, another distinct SNP enrichment QTL region, *qGI/GP/GR-D03-3*, was detected for the GI, GP, and GR across multiple environments, which could explain the relatively high PVE of 6.65–8.43%, indicating that a major gene in this genomic interval may improve seed germination (Table 3). Interestingly, 12 associated SNPs were located within the most significant haplotype block, which was almost 920 kb long and contained five haplotypes (Figures 5A, B). A haplotype analysis revealed that *qGI/GP/GR-D03-3* had two major haplotypes according to SNP location. Comparatively, Hap1 had a higher GP than Hap1 (Figures 5C, D). In total, 46 candidate genes contained in the *qGI/GP/GR-D03-3* region on chromosome D03 were identified. Among them, *Ghir_D03G009280* was annotated as auxin response factor 9 (*ARF9*) in *Arabidopsis* (Supplementary Table S6), and its homologs played a crucial role in seed dormancy. The RNA-seq and qRT-PCR assays also showed that *Ghir_D03G009280* had higher expression levels during the seed germination stage, suggesting a positive regulatory effect (Figure 5E; Supplementary Figure S5).

10.3389/fpls.2023.1254365

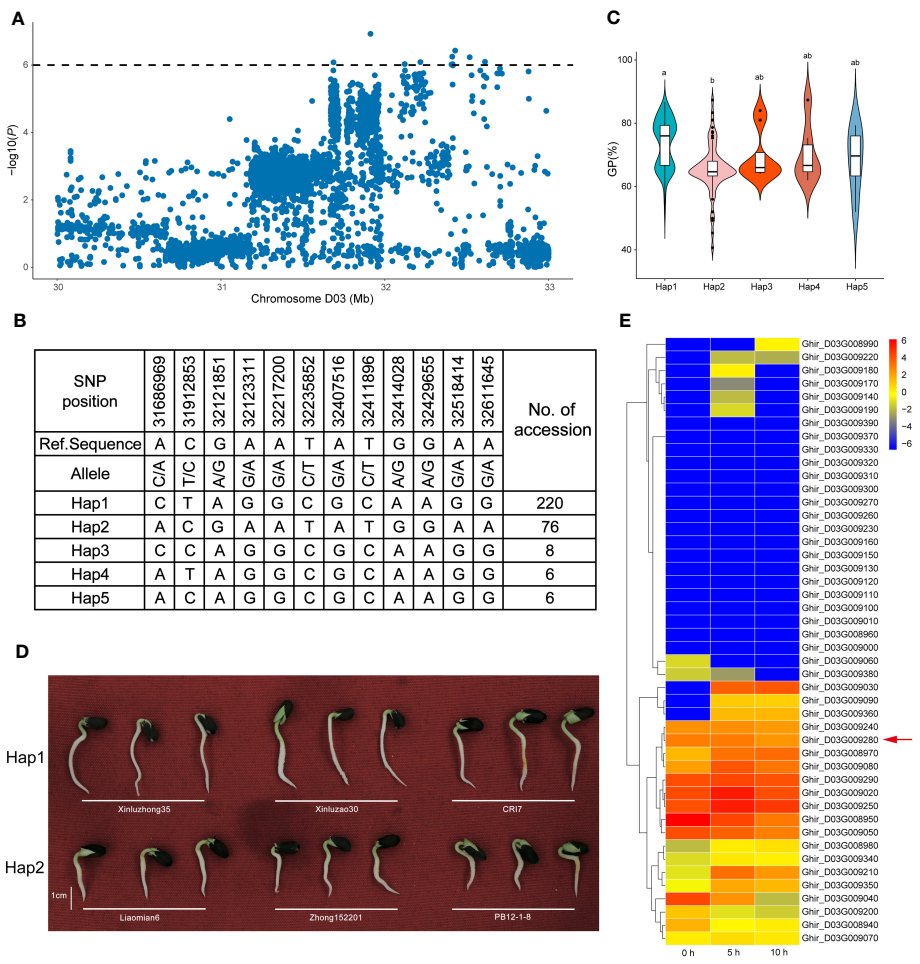
TABLE 3 Significant quantitative trait locus (QTLs) associated with seed vigor-related traits.

QTLs	SNP	Chromosome	Position (bp)	Trait	Environment	Allele	-log$_{10}$ (*P*)	Phenotypic variation explained (%)
qGR-A09-1	rsA09_7745467	A09	7,745,467	GR	BLUP; E1	T/C	6.19	6.76
	rsA09_7791621		7,791,621	GR	BLUP; E1	A/G	6.25	6.82
	rsA09_7878527		7,878,527	GR	BLUP; E1; E2	T/C	7.74	8.56
	rsA09_7908017		7,908,017	GR	BLUP; E1	A/G	6.20	6.76
	rsA09_7954329		7,954,329	GR	BLUP; E1; E2	G/C	6.88	7.55
	rsA09_7954353		7,954,353	GR	BLUP; E1; E2	A/G	6.88	7.55
	rsA09_7962794		7,962,794	GR	BLUP; E1; E2	G/A	7.22	7.95
qGP-A10-1	rsA10_112752002	A10	112,752,002	GP	BLUP; E2; E3	C/T	7.39	8.15
qGR/GI-D03-1	rsD03_15149331	D03	15,149,331	GI	E1	C/T	6.74	7.39
				GR	E1; E2	C/T	6.07	6.61
	rsD03_15180622	D03	15,180,622	GI	E1	T/C	6.08	6.62
				GR	E2	T/C	6.24	6.81
qGI/GR-D03-2	rsD03_16442805	D03	16,442,805	GR	BLUP; E2	T/A	6.25	6.81
	rsD03_17044820	D03	17,044,820	GI	E1	A/G	6.16	6.72
				GR	E2	A/G	6.40	6.99
	rsD03_17639861	D03	17,639,861	GI	E1	A/T	7.08	7.79
				GR	E2	A/T	6.49	7.10
qGI/GP/GR-D03-3	rsD03_31686969	D03	31,686,969	GP	BLUP; E1	A/C	6.61	7.24
	rsD03_31912853	D03	31,912,853	GI	BLUP; E1	C/T	7.64	8.43
				GP	BLUP; E1	C/T	7.40	8.16
	rsD03_32121851	_D03	32,121,851	GP	BLUP; E1	G/A	6.85	7.52
	rsD03_32123311	D03	32,123,311	GP	E1	A/G	7.12	7.84
				GR	E3	A/G	6.39	6.98
	rsD03_32217200	D03	32,217,200	GP	BLUP; E1	A/G	7.03	7.72
	rsD03_32235852	D03	32,235,852	GP	E1	T/C	6.83	7.49
				GR	E3	T/C	6.10	6.65
	rsD03_32407516	D03	32,407,516	GP	BLUP; E1	A/G	6.64	7.28
	rsD03_32411896	D03	32,411,896	GP	BLUP; E1	T/C	7.04	7.74
	rsD03_32414028	D03	32,414,028	GP	BLUP; E1	G/A	6.67	7.31
	rsD03_32429655	D03	32,429,655	GP	BLUP; E1	G/A	7.31	8.05
	rsD03_32518414	D03	32,518,414	GP	BLUP; E1	A/G	7.12	7.83
	rsD03_32611645	D03	32,611,645	GP	BLUP; E1	A/G	6.86	7.53
qGI/GP/GR-D03-4	rsD03_36696073	D03	36,696,073	GI	E1	A/C	6.07	6.61
				GP	E1	A/C	8.04	8.90
				GR	BLUP; E2; E3	A/C	6.53	7.15

Discussion

The importance of seed vigor for field production

SV is an indispensable indicator of seed quality, which directly affects the rapid and uniform germination of seeds and the robust growth of seedlings and affects the tolerance of plants to abiotic stress in the early stage of seedling growth (Qun et al., 2007; Fujino et al., 2008). In recent years, mechanical DS of cotton has been widely used due to its cost-saving and labor-saving advantages, leading to rapid and uniform seed germination becoming necessary conditions for high yield and mechanization in the cotton industry. However, seeds with low SV make it difficult for mechanical DS to achieve full seeding, which leads to problems such as subsequent filling of the gaps with seedlings and final singling of seedlings (Xie et al., 2014)—for example, Xinjiang Province is the major cotton-growing area in China and experiences serious saline–alkali stress (He et al., 2023). A high SV of cotton varieties will improve seed germination in the field and thus increase the yield. In addition, cotton breeding without plastic film in Xinjiang Province to eliminate "white pollution" has become popular. The germination rate and seedling emergence rate of seeds have higher requirements for cotton without plastic film (CWPF). CWPF needs to quickly establish robust seedlings after seed germination to resist the invasion of diseases, insect pests, adverse environments, and other factors in the field. Importantly, SV is the result of genetic and environmental factors and is thus often difficult to effectively select in conventional breeding (Dai et al., 2022). This study utilized high-throughput sequencing to generate widely distributed SNP markers that cover the whole genome (Figure 1), and over 200,000,000 high-quality SNPs were detected in a diverse set of 355 cotton accessions. Combining phenotype data from multiple environments for GWAS analysis can be used to effectively identify genetic loci and candidate genes that improve SV in upland cotton, providing an effective way to improve cotton yield in Xinjiang when using the MAS method.



SNP position	31686969	31912853	32121851	32123311	32217200	32235852	32407516	32411896	32414028	32429655	32518414	32611645	No. of accession
Ref.Sequence	A	C	G	A	A	T	A	T	G	G	A	A	
Allele	C/A	T/C	A/G	G/A	G/A	C/T	G/A	C/T	A/G	A/G	G/A	G/A	
Hap1	C	T	A	G	G	C	G	C	A	A	G	G	220
Hap2	A	C	G	A	A	T	A	T	G	G	A	A	76
Hap3	C	C	A	G	G	C	G	C	A	A	G	G	8
Hap4	A	T	A	G	G	C	G	C	A	A	G	G	6
Hap5	A	C	A	G	G	C	G	C	A	A	G	G	6

FIGURE 5
Variation analysis of seed vigor (SV)-related traits associated with qGI/GP/GR-D03-3 in the candidate region. **(A)** Local Manhattan plots for SV-related genes on chromosome D03 from 30 to 33 Mb. **(B)** Top two haplotypes of *qGI/GP/GR-D03-3* in 355 upland cotton accessions. **(C)** Comparison of germination potential between accessions containing Hap1, Hap2, Hap3, Hap4, and Hap5. Letters on the violin plot indicate significant differences according to one-way ANOVA (LSD test; $P < 0.05$). **(D)** Comparison of seed germination status for 3 days between Hap1 and Hap2. **(E)** Heat map of candidate gene expression patterns in the seed germination stage (0, 5, and 10 h) on chromosome D03.

Comprehensive analysis of SV-related traits at multiple environments

To ensure the accuracy of the GWAS results, phenotypic identification in multiple environments was conducted with at least three replicates per environment. The three SV-related traits (GI, GP, and GR) were measured for seeds collected from three locations: E1, E2, and E3. Among them, GR and GP did not show an absolute normal distribution, which was also found in previous studies (Dai et al., 2022; Si et al., 2022), indicating a complex genetic basis for these SV-related traits. Through phenotypic correlation analysis, it was found that there were significant positive correlations between the three traits. The GI showed a strong correlation with GR and GP (0.71 and 0.76, respectively) (Figure 2J). The highest GI was accompanied by the highest GP and GR, which is consistent with previous findings (Si et al., 2022). Furthermore, according to the measurement results for each trait, the CV of SV-related traits in upland cotton is affected by the environment (Supplementary Table S2), resulting in different variations in the seeds of each accession harvested in different planting locations and years—for example, the CV of the GI and GR in E1 showed a larger range of variation than that in E2 and E3. Previous studies have shown that the environment in the planting area has a great influence on the growth and development of seeds (Fenner, 1992). It is speculated that the E2 and E3 (Sanya City, Hainan Province) environments with tropical climates are more suitable environments for seed growth, and the performance of the seeds may be relatively stable. In contrast, the E1 environment (Huanggang City, Hubei Province) has high precipitation and temperature during the seed maturation period, which can affect the success of pollination.

Candidate genes related to SV

In the past two decades, GWAS has become a powerful and widely used tool for analyzing the genetic mechanisms underlying complex quantitative traits in crops (Tibbs Cortes et al., 2021). At present, most research on SV mainly focuses on the mechanism under stress in upland cotton (Sun et al., 2018; Yuan et al., 2019; Zheng et al., 2021), while genetic analysis of SV-related traits associated with normal seed germination is less common (Si et al., 2022). In this study, a GWAS panel was used to measure three SV-related traits of seeds harvested in three environments. In total, six significant QTLs were stably identified on three different cotton chromosomes (Table 3), including 26 SNPs. Numerous studies have reported that several pathways are involved in regulating SV in plants, such as phytohormone signaling (GA, ABA, and auxin), amino acid metabolism, and the reactive oxygen pathway, which play a crucial role in the seed germination process and have a significant effect on the molecular mechanisms related to SV (Reed et al., 2022). It has been reported that high concentrations of ABA promote dormancy and inhibit seed germination, while high concentrations of GA promote seed germination by reversing dormancy, leading to an endogenous balance of the ABA/GA ratio but not the absolute hormone contents (Finch-Savage and Leubner-Metzger, 2006; Chen H. et al., 2020). *Ghir_A09G002650* was annotated on chromosome A09, belonging to the GA-regulated family of proteins and encoding a protein containing the GASA domain, which is most closely related to the known homolog *GASA14* in *Arabidopsis*. *GASA14* regulates the increase in plant growth through GA induction and DELLA-dependent signal transduction, which could increase resistance to abiotic stress by reducing the accumulation of ROS (Sun et al., 2013). Thus, it is speculated that *Ghir_A09G002650* has the potential to improve the SV of cotton under stress. MYB-type and bHLH-type transcription factors have been reported to be involved in the regulation of seed germination signaling in plants (Penfield et al., 2005; Reyes and Chua, 2007; Kim et al., 2015; Wang X. et al., 2022; Xu et al., 2022). Specifically, *Ghir_D03G006550* is in the *qGI/GR-D03-2* region and is homologous to *MYB52*. It has been previously shown that its shared common targets with *ERF4* regulate the development of the seed coat in *Arabidopsis* (Ding et al., 2021). *Ghir_D03G010510* encoded bHLH-type family proteins in the QTL region of *qGI/GP/GR-D03-4*, sharing 35.52% sequence identity with the PIF8 protein in *Arabidopsis*, which binds to promoter regions of AtPIF6. The expression level of *AtPIF6* during seed development plays a crucial role in establishing primary seed dormancy levels (Peters et al., 2010).

Notably, *Ghir_A09G002730* and *Ghir_D03G009280* were detected in two distinct enriched regions located on chromosome A09 (*qGR-A09-1*) and chromosome D03 (*qGI/GP/GR-D03-3*) (Figure 3). Interestingly, *Ghir_A09G002730*, within the strong-LD region at 21.9 kb upstream of rsA09_7962794 and highly expressed during the development of seed germination (Figures 4A, E), encodes a PPR superfamily protein in *Arabidopsis*. *SOAR1* belongs to the PPR protein family and acts as a core negative regulator downstream of *ABAR* and upstream of *ABI5*, participating in ABA signaling regulation of seed germination and seedling growth processes (Ma et al., 2020). We also discovered that cotton accessions carrying rsA09_7962794-A with a higher GR had a much higher allele frequency for *Ghir_A09G002730* in YZRR and NIR than in YRR and NSER (Figures 4B, C). It is possible that the planting mode of seedling raising and transplanting in YZRR and mechanized planting in the NIR all employed single-seed sowing, which increased the selection frequency of rsA09_7962794-A. In addition, we compared the genetic diversity of the region on chromosome A09 containing *Ghir_A09G002730* in different breeding periods, and it was found that cultivars bred after the 2000s had lower genetic diversity than cultivars from other stages, implying that with the continuous increase in cotton SV during the breeding process, this gene was associated with artificial selection (Figure 4D). Therefore, it is reasonable to postulate that *Ghir_A09G002730* is a new candidate gene influencing SV in cotton. *Ghir_D03G009280* caught our attention based on the gene annotation of cotton. This gene encodes an auxin response factor. Recent studies have shown that *ARF16* interacts with *ABI5* and positively regulates the ABA response during seed germination (Mei et al., 2023). Furthermore, *Ghir D03G009280*, tightly linked with haplotype Hap1, showed a significant association with GP (Figure 5C), and materials carrying the Hap1 haplotype had longer roots (Figure 5D). The RNA-seq analysis showed a high expression level of this gene during seed germination (Figure 5E). From the above-mentioned results, we inferred that *Ghir_A09G002730* and

Ghir_D03G009280 were two major candidate genes that may play an important role in cotton SV.

Conclusions

In the present study, there was a total of 121 predicted candidate genes within six stable QTL regions. Furthermore, *Ghir_A09G002730* and *Ghir_D03G009280* caught our attention based on gene expression (RNA-seq and qRT-PCR), gene annotation, and haplotype analysis, which may play a key role in regulating the germination of cotton seeds. These results will enhance our understanding of the molecular–genetic regulation of SV in cotton.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: The resequencing data (PRJNA389777) of the 355 upland cotton germplasms used in this study. For RNA-seq data: The public RNA-seq data (PRJNA248163) including SRR1695160, SRR1695161, and SRR1695162 were downloaded from the NCBI (https://www.ncbi.nlm.nih.gov/bioproject/).

Author contributions

LL: Conceptualization, Software, Visualization, Writing – original draft, Writing – review & editing. YH: Data curation, Formal Analysis, Investigation, Software, Visualization, Writing – original draft. YW: Methodology, Writing – review & editing. YY: Writing – original draft. RL: Software, Visualization, Writing – review & editing. JW: Visualization, Writing – review & editing. MY: Methodology, Supervision, Writing – review & editing. SZ: Conceptualization, Investigation, Software, Writing – review & editing. FZ: Writing – review & editing. JH: Conceptualization, Writing – review & editing. SY: Conceptualization, Writing – review & editing. ZF: Conceptualization, Investigation, Supervision, Writing – review & editing.

Funding

This research was sponsored by the National Key Laboratory of Cotton Bio-breeding and Integrated Utilization Open Fund (CB2023A09).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1254365/full#supplementary-material

References

Andrews, S. (2010). "FastQC: a quality control tool for high throughput sequence data," *Babraham Bioinformatics* (United Kingdom: Babraham Bioinformatics, Babraham Institute, Cambridge).

Chen, L., Liu, L., Lu, B., Ma, T., Jiang, D., Li, J., et al. (2020). Exogenous melatonin promotes seed germination and osmotic regulation under salt stress in cotton (*Gossypium hirsutum* L.). *PloS One* 15, e0228241. doi: 10.1371/journal.pone.0228241

Chen, H., Ruan, J., Chu, P., Fu, W., Liang, Z., Li, Y., et al. (2020). AtPER1 enhances primary seed dormancy and reduces seed germination by suppressing the ABA catabolism and GA biosynthesis in Arabidopsis seeds. *Plant J.* 101, 310–323. doi: 10.1111/tpj.14542

Chen, Z. J., Scheffler, B. E., Dennis, E., Triplett, B. A., Zhang, T., Guo, W., et al. (2007). Toward sequencing cotton (*Gossypium*) genomes. *Plant Physiol.* 145, 1303–1310. doi: 10.1104/pp.107.107672

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Dai, L., Lu, X., Shen, L., Guo, L., Zhang, G., Gao, Z., et al. (2022). Genome-wide association study reveals novel QTLs and candidate genes for seed vigor in rice. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1005203

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10, giab008. doi: 10.1093/gigascience/giab008

Dimaano, N. G. B., Ali, J., Mahender, A., Sta. Cruz, P. C., Baltazar, A. M., Diaz, M. G. Q., et al. (2020). Identification of quantitative trait loci governing early germination and seedling vigor traits related to weed competitive ability in rice. *Euphytica* 216, 159. doi: 10.1007/s10681-020-02694-8

Ding, A., Tang, X., Yang, D., Wang, M., Ren, A., Xu, Z., et al. (2021). ERF4 and MYB52 transcription factors play antagonistic roles in regulating homogalacturonan de-methylesterification in Arabidopsis seed coat mucilage. *Plant Cell* 33, 381–403. doi: 10.1093/plcell/koaa031

Feng, Z., Li, L., Tang, M., Liu, Q., Ji, Z., Sun, D., et al. (2022). Detection of stable elite haplotypes and potential candidate genes of boll weight across multiple environments via GWAS in upland cotton. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.929168

Fenner, M. (1992). Environmental influences on seed size and composition. *Hortic. Rev.* 13, 183–213. doi: 10.1002/9780470650509.ch5

Finch-Savage, W. E., and Leubner-Metzger, G. (2006). Seed dormancy and the control of germination. *New Phytol.* 171, 501–523. doi: 10.1111/j.1469-8137.2006.01787.x

Fujino, K., Sekiguchi, H., Matsuda, Y., Sugimoto, K., Ono, K., and Yano, M. (2008). Molecular identification of a major quantitative trait locus, qLTG3–1, controlling low-temperature germinability in rice. *Proc. Natl. Acad. Sci. U.S.A.* 105, 12623–12628. doi: 10.1073/pnas.0805303105

Fujino, K., Sekiguchi, H., Sato, T., Kiuchi, H., Nonoue, Y., Takeuchi, Y., et al. (2004). Mapping of quantitative trait loci controlling low-temperature germinability in rice (Oryza sativa L.). *Theor. Appl. Genet.* 108, 794–799. doi: 10.1007/s00122-003-1509-4

Gu, Q., Ke, H., Liu, C., Lv, X., Sun, Z., Liu, Z., et al. (2021). A stable QTL qSalt-A04-1 contributes to salt tolerance in the cotton seed germination stage. *Theor. Appl. Genet.* 134, 2399–2410. doi: 10.1007/s00122-021-03831-0

Guo, A., Su, Y., Nie, H., Li, B., Ma, X., and Hua, J. (2022). Identification of candidate genes involved in salt stress response at germination and seedling stages by QTL mapping in upland cotton. *G3 (Bethesda)* 12, jkac099. doi: 10.1093/g3journal/jkac099

He, Y., Cheng, J., He, Y., Yang, B., Cheng, Y., Yang, C., et al. (2019a). Influence of isopropylmalate synthase Os IPMS 1 on seed vigour associated with amino acid and energy metabolism in rice. *Plant Biotechnol. J.* 17, 322–337. doi: 10.1111/pbi.12979

He, P., Li, J., Yu, S. E., Ma, T., Ding, J., Zhang, F., et al. (2023). Soil moisture regulation under mulched drip irrigation influences the soil salt distribution and growth of cotton in Southern Xinjiang, China. *Plants* 12, 791. doi: 10.3390/plants12040791

He, Y., Yang, B., He, Y., Zhan, C., Cheng, Y., Zhang, J., et al. (2019b). A quantitative trait locus, qSE 3, promotes seed germination and seedling establishment under salinity stress in rice. *Plant J.* 97, 1089–1104. doi: 10.1111/tpj.14181

Iglesias-Fernandez, R., and Matilla, A. (2009). After-ripening alters the gene expression pattern of oxidases involved in the ethylene and gibberellin pathways during early imbibition of Sisymbrium officinale L. seeds. *J. Exp. Bot.* 60, 1645–1661. doi: 10.1093/jxb/erp029

Ishibashi, Y., Yamamoto, K., Tawaratsumida, T., Yuasa, T., and Iwaya-Inoue, M. (2008). Hydrogen peroxide scavenging regulates germination ability during wheat (Triticum aestivum L.) seed maturation. *Plant Signal Behav.* 3, 183–188. doi: 10.4161/psb.3.3.5540

Jiang, S., Yang, C., Xu, Q., Wang, L., Yang, X., Song, X., et al. (2020). Genetic dissection of germinability under low temperature by building a resequencing linkage map in japonica rice. *Int. J. Mol. Sci.* 21, 1284. doi: 10.3390/ijms21041284

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., Mcginnis, S., and Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5–W9. doi: 10.1093/nar/gkn201

Kim, T.-H., Böhmer, M., Hu, H., Nishimura, N., and Schroeder, J. I. (2010). Guard cell signal transduction network: advances in understanding abscisic acid, CO2, and Ca2+ signaling. *Annu. Rev. Plant Biol.* 61, 561–591. doi: 10.1146/annurev-arplant-042809-112226

Kim, J. H., Hyun, W. Y., Nguyen, H. N., Jeong, C. Y., Xiong, L., Hong, S. W., et al. (2015). AtMyb7, a subgroup 4 R2R3 Myb, negatively regulates ABA-induced inhibition of seed germination by blocking the expression of the bZIP transcription factor ABI 5. *Plant Cell Environ.* 38, 559–571. doi: 10.1111/pce.12415

Kolde, R. (2012). Pheatmap: pretty heatmaps. *R Package version* 1, 726.

Leymarie, J., Vitkauskaité, G., Hoang, H. H., Gendreau, E., Chazoule, V., Meimoun, P., et al. (2012). Role of reactive oxygen species in the regulation of Arabidopsis seed dormancy. *Plant Cell Physiol.* 53, 96–106. doi: 10.1093/pcp/pcr129

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint*. arXiv, 1303.3997. doi: 10.48550/arXiv.1303.3997

Li, W., Niu, Y., Zheng, Y., and Wang, Z. (2022). Advances in the understanding of reactive oxygen species-dependent regulation on seed dormancy, germination, and deterioration in crops. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.826809

Li, W., Yang, B., Xu, J., Peng, L., Sun, S., Huang, Z., et al. (2021). A genome-wide association study reveals that the 2-oxoglutarate/malate translocator mediates seed vigor in rice. *Plant J.* 108, 478–491. doi: 10.1111/tpj.15455

Li, L., Zhang, C., Huang, J., Liu, Q., Wei, H., Wang, H., et al. (2021). Genomic analyses reveal the genetic basis of early maturity and identification of loci and candidate genes in upland cotton (Gossypium hirsutum L.). *Plant Biotechnol. J.* 19, 109–123. doi: 10.1111/pbi.13446

Li, L., Zhao, S., Su, J., Fan, S., Pang, C., Wei, H., et al. (2017). High-density genetic linkage map construction by F2 populations and QTL analysis of early-maturity traits in upland cotton (Gossypium hirsutum L.). *PloS One* 12, e0182918. doi: 10.1371/journal.pone.0182918

Liu, H., Hussain, S., Zheng, M., Peng, S., Huang, J., Cui, K., et al. (2015). Dry direct-seeded rice as an alternative to transplanted-flooded rice in Central China. *Agron. Sustain. Dev.* 35, 285–294. doi: 10.1007/s13593-014-0239-0

Ma, Y., Zhang, S., Bi, C., Mei, C., Jiang, S.-C., Wang, X.-F., et al. (2020). Arabidopsis exoribonuclease USB1 interacts with the PPR-domain protein SOAR1 to negatively regulate abscisic acid signaling. *J. Exp. Bot.* 71, 5837–5851. doi: 10.1093/jxb/eraa315

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Mei, S., Zhang, M., Ye, J., Du, J., Jiang, Y., and Hu, Y. (2023). Auxin contributes to jasmonate-mediated regulation of abscisic acid signaling during seed germination in Arabidopsis. *Plant Cell* 35, 1110–1133. doi: 10.1093/plcell/koac362

Penfield, S., Josse, E.-M., Kannangara, R., Gilday, A. D., Halliday, K. J., and Graham, I. A. (2005). Cold and light control seed germination through the bHLH transcription factor SPATULA. *Curr. Biol.* 15, 1998–2006. doi: 10.1016/j.cub.2005.11.010

Peng, L., Sun, S., Yang, B., Zhao, J., Li, W., Huang, Z., et al. (2022). Genome-wide association study reveals that the cupin domain protein OsCDP3. 10 regulates seed vigour in rice. *Plant Biotechnol. J.* 20, 485–498. doi: 10.1111/pbi.13731

Peters, S., Egert, A., Stieger, B., and Keller, F. (2010). Functional identification of Arabidopsis ATSIP2 (At3g57520) as an alkaline α-galactosidase with a substrate specificity for raffinose and an apparent sink-specific expression pattern. *Plant Cell Physiol.* 51, 1815–1819. doi: 10.1093/pcp/pcq127

Qun, S., Wang, J.-H., and Sun, B.-Q. (2007). Advances on seed vigor physiological and genetic mechanisms. *Agric. Sci. China* 6, 1060–1066. doi: 10.1016/S1671-2927(07)60147-3

Reed, R. C., Bradford, K. J., and Khanday, I. (2022). Seed germination and vigor: ensuring crop sustainability in a changing climate. *Heredity* 128, 450–459. doi: 10.1038/s41437-022-00497-2

Reyes, J. L., and Chua, N. H. (2007). ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination. *Plant J.* 49, 592–606. doi: 10.1111/j.1365-313X.2006.02980.x

Ryu, H., and Cho, Y.-G. (2015). Plant hormones in salt stress tolerance. *J. Plant Biol.* 58, 147–155. doi: 10.1007/s12374-015-0103-z

Sawan, Z. M. (2016). Cottonseed yield and its quality as affected by mineral nutrients and plant growth retardants. *Cogent Biol.* 2, 1245938. doi: 10.1080/23312025.2016.1245938

Shi, H., Guan, W., Shi, Y., Wang, S., Fan, H., Yang, J., et al. (2020). QTL mapping and candidate gene analysis of seed vigor-related traits during artificial aging in wheat (Triticum aestivum). *Sci. Rep.* 10, 1–13. doi: 10.1038/s41598-020-75778-z

Shikha, K., Shahi, J., Vinayan, M., Zaidi, P., Singh, A., and Sinha, B. (2021). Genome-wide association mapping in maize: status and prospects. *3 Biotech.* 11, 244. doi: 10.1007/s13205-021-02799-4

Shin, J.-H., Blay, S., Mcneney, B., and Graham, J. (2006). LDheatmap: an R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *J. Stat. Softw.* 16, 1–9. doi: 10.18637/jss.v016.c03

Si, A., Sun, Z., Li, Z., Chen, B., Gu, Q., Zhang, Y., et al. (2022). A genome wide association study revealed key single nucleotide polymorphisms/genes associated with seed germination in Gossypium hirsutum L. *Front. Plant Sci.* 13, 844946. doi: 10.3389/fpls.2022.844946

Su, J., Fan, S., Li, L., Wei, H., Wang, C., Wang, H., et al. (2016a). Detection of favorable QTL alleles and candidate genes for lint percentage by GWAS in Chinese upland cotton. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2022.844946

Su, J., Li, L., Pang, C., Wei, H., Wang, C., Song, M., et al. (2016b). Two genomic regions associated with fiber quality traits in Chinese upland cotton under apparent breeding selection. *Sci. Rep.* 6, 1–14. doi: 10.1038/srep38496

Su, J., Li, L., Zhang, C., Wang, C., Gu, L., Wang, H., et al. (2018). Genome-wide association study identified genetic variations and candidate genes for plant architecture component traits in Chinese upland cotton. *Theor. Appl. Genet.* 131, 1299–1314. doi: 10.1007/s00122-018-3079-5

Sun, Z., Li, H., Zhang, Y., Li, Z., Ke, H., Wu, L., et al. (2018). Identification of SNPs and candidate genes associated with salt tolerance at the seedling stage in cotton (*Gossypium hirsutum* L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01011

Sun, S., Wang, H., Yu, H., Zhong, C., Zhang, X., Peng, J., et al. (2013). GASA14 regulates leaf expansion and abiotic stress resistance by modulating reactive oxygen species accumulation. *J. Exp. Bot.* 64, 1637–1647. doi: 10.1093/jxb/ert021

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14, e20077. doi: 10.1002/tpg2.20077

Veisi, S., Sabouri, A., and Abedi, A. (2022). Meta-analysis of QTLs and candidate genes associated with seed germination in rice (*Oryza sativa* L.). *Physiol. Mol. Biol. Plants* 28, 1587–1605. doi: 10.1007/s12298-022-01232-1

Vogt, F., Shirsekar, G., and Weigel, D. (2022). vcf2gwas: Python API for comprehensive GWAS analysis using GEMMA. *Bioinformatics* 38, 839–840. doi: 10.1093/bioinformatics/btab710

Wang, L., Li, J., Yang, F., Dai, D., Li, X., and Sheng, Y. (2022). A preliminary mapping of QTL qsg5. 1 controlling seed germination in melon (Cucumis melo L.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.925081

Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., et al. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.* 51, 224–229. doi: 10.1038/s41588-018-0282-x

Wang, X., Wu, R., Shen, T., Li, Z., Li, C., Wu, B., et al. (2022). An R2R3-MYB transcription factor OsMYBAS1 promotes seed germination under different sowing depths in transgenic rice. *Plants* 11, 139. doi: 10.3390/plants11010139

Wang, X., Zou, B., Shao, Q., Cui, Y., Lu, S., Zhang, Y., et al. (2018). Natural variation reveals that OsSAP16 controls low-temperature germination in rice. *J. Exp. Bot.* 69, 413–421. doi: 10.1093/jxb/erx413

Wickham, H. (2011). ggplot2. Wiley interdisciplinary reviews: computational statistics *Wiley interdisciplinary reviews: computationalstatistics* 3, 180–185. doi: 10.1002/wics.147

Xie, L., Tan, Z., Zhou, Y., Xu, R., Feng, L., Xing, Y., et al. (2014). Identification and fine mapping of quantitative trait loci for seed vigor in germination and seedling establishment in rice. *J. Integr. Plant Biol.* 56, 749–759. doi: 10.1111/jipb.12190

Xu, F., Tang, J., Wang, S., Cheng, X., Wang, H., Ou, S., et al. (2022). Antagonistic control of seed dormancy in rice by two bHLH transcription factors. *Nat. Genet.* 54, 1972–1982. doi: 10.1038/s41588-022-01240-7

Yamaguchi, S. (2008). Gibberellin metabolism and its regulation. *Annu. Rev. Plant Biol.* 59, 225–251. doi: 10.1146/annurev.arplant.59.032607.092804

Yamauchi, Y., Ogawa, M., Kuwahara, A., Hanada, A., Kamiya, Y., and Yamaguchi, S. (2004). Activation of gibberellin biosynthesis and response pathways by low temperature during imbibition of Arabidopsis thaliana seeds. *Plant Cell* 16, 367–378. doi: 10.1105/tpc.018143

Ye, N., Zhu, G., Liu, Y., Zhang, A., Li, Y., Liu, R., et al. (2012). Ascorbic acid and reactive oxygen species are involved in the inhibition of seed germination by abscisic acid in rice seeds. *J. Exp. Bot.* 63, 1809–1822. doi: 10.1093/jxb/err336

Yuan, Y., Xing, H., Zeng, W., Xu, J., Mao, L., Wang, L., et al. (2019). Genome-wide association and differential expression analysis of salt tolerance in *Gossypium hirsutum* L at the germination stage. *BMC Plant Biol.* 19, 1–19. doi: 10.1186/s12870-019-1989-2

Zhang, C., Li, L., Liu, Q., Gu, L., Huang, J., Wei, H., et al. (2019). Identification of loci and candidate genes responsible for fiber length in upland cotton (*Gossypium hirsutum* L.) via association mapping and linkage analyses. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00053

Zheng, J., Zhang, Z., Gong, Z., Liang, Y., Sang, Z., Xu, Y., et al. (2021). Genome-wide association analysis of salt-tolerant traits in terrestrial cotton at seedling stage. *Plants* 11, 97. doi: 10.3390/plants11010097

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1. doi: 10.3835/plantgenome2008.02.0089

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 05 September 2023
DOI 10.3389/fpls.2023.1247014

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Parameswaran C,
ICAR-National Rice Research Institute, India
Md Shamim,
Bihar Agricultural University, India

*CORRESPONDENCE
Sung-Ryul Kim
s.r.kim@irri.org

RECEIVED 25 June 2023
ACCEPTED 14 August 2023
PUBLISHED 05 September 2023

CITATION
Simon EV, Hechanova SL, Hernandez JE, Li C-P, Tülek A, Ahn E-K, Jairin J, Choi I-R, Sundaram RM, Jena KK and Kim S-R (2023) Available cloned genes and markers for genetic improvement of biotic stress resistance in rice.
Front. Plant Sci. 14:1247014.
doi: 10.3389/fpls.2023.1247014

Available cloned genes and markers for genetic improvement of biotic stress resistance in rice

Eliza Vie Simon[1,2], Sherry Lou Hechanova[1], Jose E. Hernandez[2], Charng-Pei Li[3], Adnan Tülek[4], Eok-Keun Ahn[5], Jirapong Jairin[6], Il-Ryong Choi[1,5], Raman M. Sundaram[7], Kshirod K. Jena[8] and Sung-Ryul Kim[1*]

[1]Rice Breeding Innovation Department, International Rice Research Institute (IRRI), Laguna, Philippines, [2]Institute of Crop Science (ICropS), University of the Philippines Los Baños, Laguna, Philippines, [3]Taiwan Agricultural Research Institute (TARI), Council of Agriculture, Taiwan, [4]Trakya Agricultural Research Institute, Edirne, Türkiye, [5]National Institute of Crop Science, Rural Development Administration (RDA), Republic of Korea, [6]Division of Rice Research and Development, Rice Department, Bangkok, Thailand, [7]ICAR-Indian Institute of Rice Research, Rajendranagar, Hyderabad, India, [8]School of Biotechnology, KIIT Deemed University, Bhubaneswar, Odisha, India

Biotic stress is one of the major threats to stable rice production. Climate change affects the shifting of pest outbreaks in time and space. Genetic improvement of biotic stress resistance in rice is a cost-effective and environment-friendly way to control diseases and pests compared to other methods such as chemical spraying. Fast deployment of the available and suitable genes/alleles in local elite varieties through marker-assisted selection (MAS) is crucial for stable high-yield rice production. In this review, we focused on consolidating all the available cloned genes/alleles conferring resistance against rice pathogens (virus, bacteria, and fungus) and insect pests, the corresponding donor materials, and the DNA markers linked to the identified genes. To date, 48 genes (independent loci) have been cloned for only major biotic stresses: seven genes for brown planthopper (BPH), 23 for blast, 13 for bacterial blight, and five for viruses. Physical locations of the 48 genes were graphically mapped on the 12 rice chromosomes so that breeders can easily find the locations of the target genes and distances among all the biotic stress resistance genes and any other target trait genes. For efficient use of the cloned genes, we collected all the publically available DNA markers (~500 markers) linked to the identified genes. In case of no available cloned genes yet for the other biotic stresses, we provided brief information such as donor germplasm, quantitative trait loci (QTLs), and the related papers. All the information described in this review can contribute to the fast genetic improvement of biotic stress resistance in rice for stable high-yield rice production.

KEYWORDS

biotic stress, marker-assisted selection, brown planthopper, blast, bacterial blight, marker, rice

1 Introduction

Rice (*Oryza sativa* L.) is a staple food of more than 50% of the world's population; notably, it is the most important crop in Asian countries. Recently, rice consumption has been rapidly increasing in Africa as well (Seck et al., 2012). Stable high-yield production of rice is highly associated with global food security (Bandumula, 2018). However, rice plants are inevitably encountering pressing challenges from different types of biotic/abiotic stresses that cause significant rice grain yield reduction (Khush, 2005; Dixit et al., 2020). Biotic stresses caused by pests and diseases pose a significant risk to global rice yield production by 52%, of which approximately 30% of these damages are due to pathogen infection (Savary et al., 2019; Jamaloddin et al., 2021). In addition, global climate change is a major threat to global food security (Schneider and Asch, 2020). A changing climate will influence the distribution and possibly the impact of rice diseases (Bebber, 2015; Chaloner et al., 2021) as well as host and disease interactions, mechanism, reproduction, and survival of pathogens (Velásquez et al., 2018).

Rice plants are attacked by diverse biotic agents, including insect pests, fungal and bacterial pathogens, and viruses. The prevalence of species of pathogens and biotypes/pathotypes is variable based on the environmental condition and geographical locations. Over the past decades, outbreaks due to pests and diseases have caused serious economic damage to rice-growing countries from time to time, locally and globally. For instance, some devastating damage from brown planthopper (BPH) infestation has been reported in different years in many rice-growing countries, including tropical and temperate Asia (Dyck and Thomas, 1979; Jena and Kim, 2010). Rice blast disease causes a loss of rice yield sufficient to feed 60 million people worldwide (Fahad et al., 2019; Singh et al., 2020). As a viral disease, a series of large-scale outbreaks of tungro were recorded in many tropical Asian countries, and it causes yield losses of 5% to 10% annually (Dai and Beachy, 2009). In Africa, rice yellow mottle virus (RYMV) is one of the most problematic biotic stresses, it reduces grain yield by 10%–100%, and severe attacks can lead to plant death (Kouassi et al., 2005). Still, today, severe biotic stress damage is reported in local or national media, implying that biotic stress damage affects local rice farmers, particularly small and marginal farmers.

There are several practical methods used to control pathogens, such as chemical spraying, crop rotation, field management, and host resistance. Among these, genetic improvement of host resistance by introgression of resistance genes through breeding and cultivation of resistant varieties is the most cost-effective and environmental-friendly strategy for controlling biotic agents. Thus, much effort has been exerted by scientists and breeders in isolating germplasms possessing resistance to a variety of biotic stresses from cultivars, landraces, and wild rice species in the genus *Oryza*. Through genetic analysis, they have also identified the genetic factors (quantitative trait loci (QTLs)/genes) that provide resistance from the isolated germplasm.

Once the genetic factors conferring biotic stress resistance are identified, they can be easily and effectively transferred to the target background varieties by marker-assisted selection (MAS) compared to the conventional phenotype-based selection. DNA markers that can discriminate the alleles (sequences) between the donor and elite susceptible variety play important roles in efficiently deploying the identified genetic factors. Different types of molecular markers have been developed based on the types of sequence variations (short or long InDels and single-nucleotide polymorphisms (SNPs)) and successfully applied in the genetics and breeding of rice. Among them, the PCR-gel-based markers such as simple sequence repeat (SSR) markers, also called rice microsatellite (RM) markers, InDel markers, dominant PCR markers, tetra-primer method markers, and cleaved amplified polymorphic sequence (CAPS: PCR-restriction enzyme application-gel) markers are the most common in rice MAS breeding due to simplicity, in-house accessibility, and easiness to breeders (McCouch et al., 2002; Chen H, et al., 2011; Wang et al., 2012; Kim et al., 2016; Nadeem et al., 2018).

To improve the genetic potential of biotic stress resistance through MAS, two key factors are essential: genetic factors (QTLs and genes) and molecular tools (DNA markers). Compared to the QTL level of genetic factors, the cloned genes/alleles have some advantages: i) the genetic effect will be quite reliable because it was functionally validated by using transgenic approaches such as complementation test, RNAi, and CRISPR tools; ii) the exact physical location of the gene is identified, and thus, it enables a precision marker-assisted introgression of the target gene without linkage drag caused by the neighboring genes. Many biotic stress resistance genes were cloned from cultivars, landraces, and wild rice germplasm possessing "natural variations", but some of the genes were identified by transgenic approaches such as overexpression, RNAi, and CRISPR and also by using rice T-DNA tagging lines. Several review papers already covered recent advances in understanding the molecular mechanism of biotic stress resistances for BPH (Yan et al., 2023), blast (Liu W, et al., 2013; Li et al., 2019), and bacterial blight (Jiang et al., 2020; Pradhan et al., 2020) and also broad-spectrum disease resistance in rice (Ke et al., 2017; Liu et al., 2021). In this review, we focused on consolidating all the available cloned genes/alleles with corresponding donors possessing "natural variations" and all the related DNA markers for the breeding aspects. In addition, we briefly described some review papers and recent publications about the QTLs or germplasm if the cloned genes are not available for specific pathogens. We aimed to provide breeding-related information so that breeders can easily select the available resistant genes/alleles and the associated markers for the fast deployment of the proper genes/alleles in their breeding programs to deal with stable high-yield rice production and climate change.

2 Precision marker-assisted breeding by using the cloned genes/alleles

Deployment of QTLs and genes through marker-assisted breeding has been successfully improving the genetic potential of target traits in many crops. However, occasional acquisition of biotic stress resistance by the breeding process used to be associated

with yield penalties in crops (Brown, 2002) and also grain quality in rice (Fukuoka et al., 2009) probably due to the presence of unfavorable genes located in the vicinity of the target biotic stress resistance locus (also called linkage drag). Thus, precise introgression of biotic stress resistance genes through marker-assisted breeding of the cloned genes can reduce unexpected penalties in yield, grain qualities, and also other agronomic traits in the final breeding products. Recent advances in DNA sequencing, genotyping technologies, genome-wide association study (GWAS), functional genomics, and gene validation by using transgenic approaches have been accelerating the identification of the causal genes governing the target traits. Notably, many biotic stress resistance genes from the previously identified major QTLs have been gradually cloned. The cloned genes/alleles possessing natural variations are valuable for the genetic improvement of biotic stress resistance in rice. Furthermore, unlike QTL level genetic factors (more than several hundred kb), breeders can precisely introgress the gene (100 kb) using marker-based recombinant selection to avoid unwanted phenotypes caused by linkage drag in the final breeding lines because the exact physical location of the causal gene is clearly known. To date, 48 genes have been cloned for the major rice biotic stress, including bacterial blight (BB), blast, BPH, and rice viruses. The cloned gene names, gene IDs of rice databases (RAP-DB and MSU), encoding proteins, the physical location of the genes, donor germplasm, and its original research papers are summarized in this review. In some cases, the previously reported major QTLs from different sources were identified as the same gene (same locus) with different alleles (different sequences). For example, *BPH1=BPH10=BPH18=BPH21/BPH2=BPH26/BHP7/BPH9* on the long arm of Chr 12 ("=" and "/" means identical and different alleles, respectively) and *Pi9/Pi2/Piz-t/Pi50/PigmR* on the short arm of Chr 6 are the different resistant alleles but the same locus. Due to the same physical locations, those alleles cannot be pyramided, and thus, the potential best allele should be selected and used in the breeding program. In this review, we focused on the cloned biotic stress resistance genes with the gene-linked markers. Moreover, we briefly mentioned some genetic resources such as QTLs or donor materials if there are no cloned genes yet for some biotic stresses.

3 Insect pests and available genetic resources

Globally, more than 100 species of insects attack rice plants, and approximately 20 of them can cause economic damage (Pathak and Khan, 1994). Major insect pests of rice are stem borers, leafhoppers and planthoppers, gall midges, and grain-sucking bugs. Efforts to isolate the resistant germplasm and genetic factors against insect pests identified a number of QTLs for the major insect pests. At the gene level, a handful of genes were cloned for only BPH resistance, but to date, no genes have been cloned yet for other insect pest resistance. Here, we described BPH resistance genes cloned and some genetic resources (QTLs and donor sources) for other insect pests.

3.1 Brown planthopper (*Nilaparvata lugens*)

Among the major insect pests, BPH is one of the most destructive pests, especially in Asian countries including both tropical and temperate zones, causing severe economic loss to the rice crop through directly sucking phloem sap, often causing "hopper burn", and it serves as a vector for transmission of rice ragged stunt virus (RRSV) and rice grassy stunt virus (RGSV) (Cabauatan et al., 2009). To date, more than 45 genetic loci providing BPH resistance have been identified from diverse plant materials, including cultivars, landraces, and wild rice species. Among them, seven genes (seven independent loci) comprising 10 different alleles for BPH resistance were cloned, that is, *BPH14*, *BPH30*, *BPH17*, *BPH6*, *BPH29*, *BPH32=BPH3*, and *BPH1=BPH10=BPH18=BPH21/BPH2=BPH26/BHP7/BPH9*. The cloned genes with physical locations, RAPDB/MSU gene ID, protein encoded, donor sources, and corresponding references are summarized in Table 1. *BPH14* gene encoding nucleotide-binding site (NBS) and leucine-rich repeats (LRRs), "NBS-LRR" or "NLR" in short, was first cloned from the previously mapped *Qbp1* on Chr 3 of the *Oryza officinalis* introgression by genetic mapping and following transgenic complementation test (Du et al., 2009). With similar approaches, the *BPH17* QTL on Chr 4S of the Sri Lankan rice variety, Rathu Heenati (Sun et al., 2005), revealed that three repeats of lectin receptor kinase gene (*OsLecRK1-OsLecRK3*) are responsible for BPH resistance (Liu et al., 2015). However, Liu et al. (2015) named the gene identified from the *BPH17* QTL as *BPH3* gene, and thus, it might cause confusion with the original *BPH3* QTL mapped on Chr 6S of donors (PTB33 and Rathu Heenati varieties) (Jairin et al., 2007). To avoid confusion, we followed the original *BPH17* QTL name as *BPH17* gene name in this review. Afterward, Ren et al. (2016) cloned the causal gene of BPH resistance from the previously fine-mapped *BPH3* locus of PTB33 (Jairin et al., 2007) using bioinformatics and transgenic validation experiments. The cloned gene encodes an unknown short consensus repeat (SCR) domain-containing protein and the *BPH3* QTL was renamed as *BPH32* (*BPH32=BPH3*) (Ren et al., 2016). Some of the BPH-resistant loci from different sources overlapped at the same locus, resulting in four clusters on chromosomes 4S, 4L, 6S, and 12L (Fujita et al., 2013; Du et al., 2020). From the largest BPH QTL cluster on Chr 12L containing *BPH1*, *BPH2*, *BPH7*, *BPH9*, *BPH10*, *BPH18*, *BPH21*, and *BPH26* (Fujita et al., 2013), *BPH26* encoding NBS-LRR protein was first cloned from the *BPH26* QTL derived from ADR52 (Tamura et al., 2014). Then, *BPH18* from the *BPH18* QTL originated from the *Oryza australiensis* introgression line (IL) (IR65482-7-216-1-2) was cloned and identified as the same gene with *BPH26* because physically two genes are located at the same locus on Chr 12L. However, the sequences, including promoter and protein-coding sequences (CDS) and also BPH reactions, were different between *BPH26* and *BPH18* (Ji et al., 2016). *BPH9* derived from Pokkali was also identified as the same gene as *BPH18*/*BPH26*, but it showed different gene sequences and also different BPH reactions (Zhao et al., 2016), suggesting that all three are the same gene (locus) but

TABLE 1 The cloned BPH resistance genes.

Gene	Chr	Location (bp)[a]	MSU_ID	RAPDB_ID	Encoding protein	Resistant/ donor allele	Inheritance pattern of R- allele	Reference
BPH14	3	35,693,286	Os03g63150	Os03g0848700	NBS-LRR	*Oryza officinalis* IL	Dominant	Du et al., 2009
BPH30	4	929,966	Os04g02520	–	Protein with two leucine-rich domains (LRDs)	AC-1613	Dominant	Shi et al., 2021
BPH17[(b)]	4	6,940,275	Os04g12540–Os04g12560–Os04g12580	Os04g0201900–Os04g0202300–Os04g0202500	A cluster of three genes encoding plasma membrane-localized lectin receptor kinases (OsLecRK1-OsLecRK3)	Rathu Heenati	Dominant	Liu et al., 2015
BPH6	4	21,396,879	Os04g35210	Os04g0431700	Atypical LRR	Swarnalata	Dominant	Guo et al., 2018
BPH29	6	484,346	Os06g01860	Os06g0107800	B3 domain-containing protein	RBPH54 (*Oryza rufipogon* IL)	Recessive	Wang Y, et al., 2015
BPH32 =*BPH3*[©]	6	1,223,069	Os06g03240	Os06g0123200	Unknown short consensus repeat (SCR) domain-containing protein	Ptb33	Dominant	Ren et al., 2016
BPH1=BPH10=BPH18=BPH21/ BPH2=BPH26/BHP7/BPH9[(d)]	12	22,886,341	Os12g37290	Os12g0559400	NBS-LRR	IR65482-7-216-1-2 (*BPH18*), ADR52 (*BPH26*), T12 (*BPH7*), Pokkali (*BPH9*)	Dominant	Tamura et al., 2014 (*BPH26*), Ji et al., 2016 (*BPH18*), Zhao et al., 2016 (*BPH9* and other alleles)

"=" means the identical allele, and "/" means the different alleles at the same locus.

[(a)]Location of the translation start codon (ATG) of the cloned genes on the rice reference genome IRGSP1.0 (https://rapdb.dna.affrc.go.jp/).

[(b)]BPH17 was identified from the mapping populations derived from the cross Rathu Heenati (R) and 02428 variety (S) by Sun et al. (2005). In a subsequent study, Liu et al. (2015) cloned the BPH resistance gene from the same materials, but the gene was probably mistakenly named BPH3 in the publication. Hence, to avoid confusion with previously reported BPH3 QTL (Jairin et al., 2007), the original name QTL name (BPH17) was given in this review.

[(c)]BPH32 was identified by using bioinformatics and transgenic gene validation experiments by Ren et al. (2016) from the previously fine-mapped BPH3 locus (Jairin et al., 2007).

[(d)]Eight BPH genes clustered on Chr 12L were identified as multi-alleles with four different sequences (four allele types) at the same locus (Zhao et al., 2016). However, the NILs with the same allele types (BPH10, BPH18, and BPH21) showed a bit different BPH resistance among the same allele types (Jena et al., 2017).

functionally different alleles. Based on the sequence analysis of the Chr 12L BPH cluster, Zhao et al. (2016) classified the eight genes into four allelotypes, *BPH1=BPH10=BPH18=BPH21/ BPH2=BPH26/BHP7/BPH9*. However, the BPH near-isogenic lines (NILs) with the same allele types (*BPH10*, *BPH18*, and *BPH21*) showed slightly different BPH resistance among the same allele types (Jena et al., 2017). Although four different functional alleles were identified on Chr 12L, they cannot be pyramided by MAS breeding due to their same locations. Guo et al. (2018) cloned the *BPH6* encoding NBS-LRR protein from the previously found *BPH6* QTL originating from the Swarnalata variety, which exhibits resistance to biotype 4, the most devastating BPH biotype in South Asia, of Bangladesh BPH populations (Kabish and Khush, 1988). The recessive gene *BPH29* located at Chr 6 was found to encode a B3-domain containing protein from the RBPH54 IL possessing BPH resistance derived from the wild rice species *Oryza rufipogon* (Wang Y, et al., 2015). *BPH30* gene located on Chr 4 of the *indica* variety AC-1613 was identified as a gene that encodes a novel protein with two leucine-rich domains (Shi et al., 2021). In addition to the cloned BPH genes, a number of QTLs and fine-mapped QTLs are also available (Fujita et al., 2013; Naik et al., 2018; Du et al., 2020). Moreover, using 10 different BPH genes/QTLs, 25 NILs possessing single or two to three genes were developed in an *indica* variety background, IR24 (Jena et al., 2017). The set of BPH NILs will be useful for screening suitable BPH genes/alleles against regional BPH biotypes and for genetic improvement of BPH

resistance in the local elite variety backgrounds. To achieve durable and broad-spectrum resistance, QTL/gene pyramiding approaches are widely used in breeding programs. Overall, the BPH-NILs with two to three genes exhibited more strong and broad-spectrum resistance than the NILs harboring a single BPH gene (Jena et al., 2017). In addition, pyramiding effects of two to three BPH gene combinations such as *BPH14* + *BPH15*, *BPH6* + *BPH12*, and *BPH13* + *BPH14* + *BPH15* were observed in different backgrounds or breeding programs (Hu et al., 2012; Qiu et al., 2012; Hu et al., 2016).

3.2 Other planthoppers

A handful of genetic factors governing resistance against planthoppers, including small brown planthopper (SBPH: *Laodelphax striatellus*), white-backed planthopper (WBPH: *Sogatella furcifera*), green leafhopper (GLH: *Nephotettix virescens*), and green rice leafhopper (GRH: *Nephotettix cincticeps*), have been identified from diverse germplasms and are well summarized in a few review papers (Fujita et al., 2013; Du et al., 2020). In this review, we only included recent progress on genetic factors to other planthoppers. A stable locus showing WBPH resistance in 2 years was found in the RM280-RM6909 region on Chr 4L from the Cheongcheong variety (Kim et al., 2021). The high resistance locus designated as *Bph38* to both BPH and WBPH was identified from *O. rufipogon* and was fine-mapped to a 79-kb region on Chr 4 (Yang et al., 2020). Phi et al. (2019) identified a major QTL (*qGRH4.2=GRH6*) conferring GRH resistance from a wild species (*Oryza nivara*_IRGC105715) and fine-mapped the locus to ~31-kb region on Chr 4. Recent studies showed a possibility that increasing resistance to multiple insects could be achieved by the pyramiding of insect resistance loci. For example, both GLH and GRH resistance was obtained by pyramiding of two GRH resistance genes, *GRH2* and *GRH4* (Horgan et al., 2018); enhanced resistance against multiple herbivore species, including zig-zag leafhopper (*Recilia dorsalis*), BPH, and WBPH, was shown by pyramiding of two to three GRH resistance loci (*GRH2* and *GRH4-6*) (Horgan et al., 2019).

3.3 Rice gall midge (*Orseolia oryzae*)

To date, 12 potential genetic factors (*Gm1–Gm12*) conferring resistance against Asian rice gall midges (*O. oryzae*) have been reported. Among them, 10, except for *Gm9* and *Gm10*, are mapped on rice chromosomes (Bentur et al., 2016; Leelagud et al., 2020). Although no *Gm* genes have been fully validated by using transgenic approaches, four QTLs were fine-mapped with potential candidate genes: *gm3* (donor: RP2068-18-3-5 breeding line from Velluthacheera) on 560-kb region of Chr4L (Sama et al., 2014), *Gm4* (donor: Abhaya) on 300-kb region of Chr 8 (Divya et al., 2015), *Gm8* (donor: Aganni) on 430-kb region of Chr 8 (Divya et al., 2018), and *gm12* (donor: MN62M) on 345-kb region of Chr 2 (Leelagud et al., 2020). These four QTLs might be useful in a breeding program. However, the donor sources showing resistance against Indian gall midge biotypes, including Velluthacheera (*gm3*), Abhaya (*Gm4*), and Aganni (*Gm8*), were susceptible to all eight Thailand gall midge populations (Leelagud et al., 2020), suggesting that the suitable genetic factors should be selected based on the potential biotypes of insects.

3.4 Other insect pests

Five QTLs associated with leaf-folder (*Cnaphalocrocis medinalis*) resistance, with 8.0%–21.1% phenotypic variance explained (PVE), were found from the double haploid population (CJ06 × TN1), and pyramiding of QTLs affected resistance to leaf-folder (Rao et al., 2010). However, reliable genetic factors controlling other insect resistance, including stem borer and grain-sucking bugs, have not been reported yet.

4 Fungal diseases and available genetic resources

Several major fungal pathogens threaten stable high-yield rice production. The major fungal diseases of rice are "bakanae disease" (pathogen: *Gibberella fujikuroi*, syn. *Fusarium fujikuroi*), "brown spot" (pathogen: *Cochliobolus miyabeanus*, syn. *Bipolaris oryzae*, *Helminthosporium oryzae*), "narrow brown leaf spot" also called "narrow brown spot" (pathogen: *Sphaerulina oryzina*, syn. *Cercospora janseana*, *Cercospora oryzae*), "false smut" (pathogen: *Ustilaginoidea virens*), "leaf scald" (pathogen: *Microdochium oryzae*), "sheath blight" (pathogen: *Rhizoctonia solani*, syn. *Thanatephorus cucumeris*), "aggregate sheath spot" (pathogen: *Rhizoctonia oryzae-sativae*), "sheath rot" (pathogen: *Sarocladium oryzae*), "stem rot" (pathogen: *Sclerotium oryzae*, syn. *Nakataea oryzae*), and "blast" (pathogen: *Magnaporthe oryzae*, syn. *Pyricularia oryzae*). Among fungal diseases, blast has been intensively studied compared to other fungal diseases. As a result, a handful of blast-resistance genes have been cloned, but no cloned genes are available yet for other fungal diseases.

4.1 Blast (pathogen: *M. oryzae*, syn. *P. oryzae*)

Among the fungal diseases, rice blast is the most devastating fungal disease of rice worldwide, causing a serious threat to the world's food security. The blast pathogen can affect all above-ground parts of a rice plant, including the leaf, collar, node, neck, parts of the panicle, and sometimes the leaf sheath (IRRI Rice Knowledge Bank). Blast disease occurs in 85 countries, and it causes a 10%–35% loss of harvest (Fisher et al., 2012), and the amount of rice damaged by blast annually is sufficient to feed 60 million people worldwide (Pennisi, 2010; Fahad et al., 2019; Singh et al., 2020). There are over 100 blast resistance QTLs/loci identified from diverse germplasm including cultivars, landraces, and wild relatives of rice (Ashkani et al., 2016; Li et al., 2019). The *Pib* (donor: *indica* cultivar Engkatek) and *Pita* (donor: *indica* cultivar Tadukan) were the first cloned blast resistance genes, and both

encode NBS-LRR domains predicted to be cytoplasmic proteins (Wang et al., 1999; Bryan et al., 2000). To date, 23 genes (23 independent loci) consisting of ~35 different alleles have been cloned, including three panicle blast resistance genes *Pb1*–*Pb3* (Table 2). The cloned genes were distributed across the rice chromosomes except for chromosomes 5, 7, and 10. Chromosomes 6 and 11 harbored four and six blast genes, respectively (*Pi9* alleles, *Pid4*, *Pid3* alleles, and *Pid2* on Chr 6; *Pia* alleles, *Pi54rh* alleles, *Pik* alleles, *Pb1*, *Pb2*, and *Pb3* on Chr 11). Several blast-resistant QTLs were identified at the same location on the short arm of Chr 6 (10.4-Mb region) from different germplasms. Finally, the causal genes were located at the NLR gene-repeated cluster (*Pi9* locus), and they are regarded as the same genes with different alleles (*Pi9*/*Pi2*=*Piz-5*/*Piz-t*/*Pi50*/*Pigm*/*Pizh*). At the *Pi9* locus, two to 13 repeats of NLR gene were laid next to each other, and the blast-resistant donors possessed nine repeats (*Pi9* and *Pi2*) or 13 repeats (*Pigm*) of NLR genes (Deng et al., 2017). There were sequence variations among the alleles of the responsive NLR gene at the *Pi9* locus, and they showed different reactions to the blast strains. In addition to the cloned genes/alleles, one major QTL (*Pi40*) was identified at the *Pi9* locus from the *O. australiensis*-derived IL (IR65482-4-136-2-2) through fine mapping (Jeung et al., 2007). The *Pi40* introgression in Korean and Turkish varieties showed resistance to a wide range of blast strains in Korea and Turkey (Jeung et al., 2007; Beser et al., 2016). Another major cluster was found on Chr 11 (25.2-Mb region) (*Pik* locus) from various donor materials, and they (*Pik*/*Pik-m*/*Pik-p*/*Pi1*/*Pi7*) were identified as allelic (Table 2). Interestingly, most of the cloned blast genes encode NBS-LRR (NLR) protein, except for four genes: *bsr-d1* (C2H2-type zinc finger protein), *pi21* (proline-rich protein), *Pid2*

TABLE 2 The cloned blast resistance genes.

Gene	Chr	Location (bp)	MSU_ID	RAPDB_ID	Encoding protein	Resistant/donor allele	Inheritance pattern of R-allele	Reference
Pit	1	2,681,220	Os01g05620	Os01g0149500	NBS-LRR	K59	Dominant	Hayashi K, et al., 2010
Pi64	1	33,098,082	Os01g57280	Os01g0781200	NBS-LRR	Yangmaogu	Dominant	Ma et al., 2015
Pi37	1	33,120,499	Os01g57310	Os01g0781700	NBS-LRR	St. No. 1	Dominant	Lin et al., 2007
Pish/*Pi35*	1	33,136,846	Os01g57340	Os01g0782100	NBS-LRR	Nipponbare (*Pish*), Hokkai 188 (*Pi35*)	Dominant	Takahashi et al., 2010 (*Pish*), Fukuoka et al., 2014 (*Pi35*)
Pib	2	35,118,769	Os02g57310	Os02g0818500	NBS-LRR	Engkatek	Dominant	Wang et al., 1999
bsr-d1	3	18,435,990	Os03g32230	Os03g0437200	C2H2-type zinc finger protein	Digu	Dominant	Li et al., 2017
pi21	4	19,835,206	Os04g32850	Os04g0401000	Proline-rich protein	Owarihatamochi	Recessive	Fukuoka et al., 2009
Pi63	4	31,554,480	Os04g52970	Os04g0620950	NBS-LRR	Kahei	Dominant	Xu et al., 2014
Pi9/*Pi2*=*Piz-5*/*Piz-t*/*Pi50*/*PigmR*$^{(e)}$/*Pizh*	6	10,387,509	Os06g17900	Os06g0286700	NBS-LRR	*Oryza minuta* IL (75-1-127) (*Pi9*), C101A51 (*Pi2*), Toride 1 (*Piz-t*), Er-Ba-Zhan (*Pi50*), Gumei 4 (*PigmR*), ZH11 (*Pizh*)	Dominant	Qu et al., 2006 (*Pi9*), Zhou et al., 2006 (*Pi2* and *Piz-t*), Su et al., 2015 (*Pi50*), Deng et al., 2017 (*PigmR*), Xie et al., 2019 (*Pizh*)
Pid4	6	10,435,819	Os06g17950	Os06g0287500	NBS-LRR	Digu	Dominant	Chen et al., 2018
Pid3/*Pi25*/*Pid3-I1*	6	13,054,818	Os06g22460	Os06g0330100	NBS-LRR	Digu (*Pid3*), Gumei2 (*Pi25*), MC276 (*Pid3-I1*)	Dominant	Shang et al., 2009 (*Pid3*), Chen J, et al., 2011 (*Pi25*), Inukai et al., 2019 (*Pid3-I1*)
Pid2	6	17,160,333	Os06g29810	Os06g0494100	B-lectin receptor kinase	Digu	Dominant	Chen et al., 2006
Pi36	8	2,878,884	Os08g05440	Os08g0150150	NBS-LRR	Kasalath (formerly known as Q61)	Dominant	Liu et al., 2007
Pi5	9	9,681,913	Os09g15840	Os09g0327600	NBS-LRR	RIL260-Moroberekan	Dominant	Lee et al., 2009
Pi56	9	9,777,527	Os09g16000	Os09g0328951	NBS-LRR	Sanhuangzhan No 2 (SHZ2)	Dominant	Liu Y, et al., 2013

(Continued)

TABLE 2 Continued

Gene	Chr	Location (bp)	MSU_ID	RAPDB_ID	Encoding protein	Resistant/donor allele	Inheritance pattern of R-allele	Reference
Pia/*Pi-CO39*	11	6,541,924	Os11g11790–Os11g11810	Os11g0225100–Os11g0225300	Two genes encoding NBS-LRR	Sasanishiki (*Pia*), CO39 (*Pi-CO39*)	Dominant	Okuyama et al., 2011 (*Pia*), Cesari et al., 2013 (*Pi-CO39*)
Pi54rh/*Pi54=Pik-h*	11	25,263,336	Os11g42010	Os11g0639100	NBS-LRR	*Oryza rhizomatis* (*Pi54rh*), Tetep (*Pi54*)	Dominant	Das et al., 2012 (*Pi54rh*), Zhang et al., 2018 (*Pi54*)
Pik/*Pik-m*/*Pik-p*/*Pi1*/*Pi7*	11	27,983,597	Os11g46200-Os11g46210	Os11g0688832–Os11g0689100	Two genes encoding NBS-LRR	Kusabue (*Pik*), Tsuyuake (*Pik-m*), K60 (*Pik-p*), C101LAC (*Pi1*), IRBL7-M (*Pi7*)	Dominant	Zhai et al., 2011 (*Pik*), Ashikawa et al., 2008 (*Pik-m*), Yuan et al., 2011 (*Pik-p*), Hua et al., 2012 (*Pi1*), Gan et al., 2010 (*Pi7*)
Pita	12	10,606,359	Os12g18360	Os12g0281300	NBS-LRR	Tadukan	Dominant	Bryan et al., 2000
Ptr = *Pita2*	12	10,822,534	Os12g18729	Os12g0285100	Armadillo repeats protein	Katy (*Ptr*), IRBLta2-Re (*Pita2*)	Dominant	Zhao et al., 2018 (*Ptr*), Meng et al., 2020 (*Pita2*)
Pb1	11	22,862,447	Os11g38580	Os11g0598500	NBS-LRR	Modan	Dominant	Hayashi N, et al., 2010
Pb2	11	27,608,621	Os11g45620	–	NBS-LRR	Jiangnanwan	ND	Yu et al., 2022
Pb3	11	27,282,232	Os11g45090	Os11g0675200	NBS-LRR	Haplotype A, Bodao	ND	Ma et al., 2022

ND, not determined.

(B-lectin receptor kinase), and *Ptr=Pita2* (armadillo repeat protein). The majority of blast-resistant donor alleles/genes are dominant except for *pi21*, which is recessive (Liu W, et al., 2013). *Pi21* encodes a proline-rich protein, and the loss-of-function allele from the resistant donor (Owarihatamochi) confers non-race-specific resistance. *pi21* gene was closely linked to the gene providing poor eating quality. However, the genes were successfully separated by recombination between two genes in the breeding lines, and blast resistance with good eating quality was achieved (Fukuoka et al., 2009). Thus, precise introgression with the cloned target genes is able to reduce the presence of unwanted phenotypes in the final breeding products caused by "linkage drag". Among the cloned blast genes, *Pi50*, *Pizh*, *Pi54rh*, *Pi56*, *Pi64*, *PigmR*, and *Ptr=Pita2* alleles were known as broad-spectrum resistance (Liu et al., 2021). A few sets of NILs with blast resistance sources were developed in both *japonica* and *indica* backgrounds: 20 NILs with 11 blast QTLs/genes in *japonica* background Lijiangxintuanheigu (LTH) (Telebanco-Yanoria et al., 2010) and 28 NILs with 14 QTLs/genes in an *indica* background, CO39 (Telebanco-Yanoria et al., 2011). Moreover, both NIL sets were tested by 20 blast isolates collected in the Philippines. Recently, 21 NILs with 18 QTLs/genes in another *indica* background, US-2, were developed, and the NILs were tested with 31 isolates from Asia (Japan, China, the Philippines, Indonesia, Vietnam, Cambodia, Bangladesh, and Laos) and Africa (Nigeria, Kenya, and Benin) (Fukuta et al., 2022). In blast bioassay with the NIL sets above, most of the genes/QTLs showed differential reactions against different isolates, even in the same country collections, suggesting that the selection of suitable blast genes/alleles based on the local pathotypes/isolates is important to develop blast resistant varieties. Among the blast genes used in the NIL development above, NIL-*Pi9* exhibited resistance or moderate resistance to all 31 isolates from Asia and Africa (Fukuta et al., 2022), suggesting that *Pi9* allele might be useful to breed blast-resistant variety across the rice cultivation countries. The sets of NILs and blast screening data against various isolates will be very useful to pathology studies, the selection of suitable genes/alleles against regional isolates, and breeding programs. To achieve durable and broad-spectrum resistance, pyramiding of resistance genes (two or more) in one background is usually used in the breeding program. There are various gene combinations of blast genes that prove the enhanced blast resistance in both *indica* and *japonica* rice against several blast isolates. Two genes–pyramided lines with *Pi37* + *Pid3*, *Pi5* + *Pi54*, *Pi54* +*Pid3*, and *Pigm* + *Pi37* exhibited significantly enhanced resistance and observable additive effects (Jiang et al., 2019). The gene combinations *Pigm* + *Pi1*, *Pigm* + *Pi54*, and *Pigm* + *Pi33* displayed broad-spectrum resistance (Wu et al., 2019). Broad-spectrum blast resistance was also achieved in the temperate *japonica* varieties by pyramiding three to four genes with *Piz*, *Pib*, *Pik*, *Pita*, and *Pita2* (Zampieri et al., 2023). As proven in many previous studies, stacking suitable blast genes/alleles has strong potential to obtain durable and broad-spectrum resistance in the breeding program.

In contrast to leaf blast resistance, genetic resources for blast disease on other organs/tissues are relatively poor. The first panicle blast resistance gene, *Pb1*, encoding NBS-LRR was cloned from an *indica* cultivar Modan (Hayashi N, et al., 2010). Afterward, it was found that panicle blast resistance by *Pb1* is dependent on at least four other loci (Inoue et al., 2017), suggesting that a level of panicle blast resistance with *Pb1* will be influenced by other genetic factors or background materials. Recently, two additional panicle blast

resistance genes, *Pb2* and *Pb3*, were identified through GWAS and validated by transgenic approaches (Ma et al., 2022; Yu et al., 2022). Both genes encode NBS-LRR proteins and are physically close to each other (~360-kb distance between *Pb2* and *Pb3*). Some of the cloned leaf blast genes, such as *Pi25* (Chen J, et al., 2011), *PigmR* (Deng et al., 2017), and *Pid4* (Chen et al., 2018), also showed some level of panicle blast resistance.

4.2 Bakanae disease (pathogen: *G. fujikuroi*, syn. *F. fujikuroi*)

To identify the genetic factors governing bakanae disease resistance, QTL mapping and GWAS have been conducted and identified a handful of QTLs on chromosomes 1, 3, 4, 9, and 10 from several different donors, but no genes have been cloned yet. Three major QTLs (*qBK1*, *qBK1.1*, and *qFfR1*) were fine-mapped on the Chr 1 region between 23.32 and 23.67 Mb (Lee et al., 2021).

4.3 False smut (pathogen: *U. virens*)

A number of QTLs for false smut resistance have been identified by QTL mapping with bi-parental populations (Andargie et al., 2018; Han et al., 2020; Neelam et al., 2022) and GWAS (Hiremath et al., 2021). The results suggested that false smut resistance seems to quantitate traits governed by multiple genes. Among the QTLs, *qFsr8–1* originated from the Chinese rice landrace MR183–2 and showed the highest PVE (26.0%).

4.4 Sheath blight (pathogen: *R. solani*, syn. *T. cucumeris*)

More than 200 QTLs associated with sheath blight (ShB) resistance have been identified from the diverse mapping populations (Zarbafi and Ham, 2019; Goad et al., 2020). Among all the identified ShB QTLs, two loci on Chr 9 (*qShB9-2*) and Chr 11 (*qSBR11-1*) contribute 25% and 14% of PVE, respectively, are the major effect QTLs (Molla et al., 2020), and may be useful in a breeding program.

4.5 Brown spot (pathogen: *C. miyabeanus*, syn. *B. oryzae*, *H. oryzae*)

For brown spot (BS) resistance, susceptible and resistant germplasms were identified by several studies. Several cultivars that have been categorized as resistant did not show complete resistance (immunity), but they showed quantitative resistance to BS. To date, more than 20 QTLs with low–mild phenotypic variation (<20%) were identified from several mapping populations, including recombinant inbred lines (RILs), doubled haploid lines (DHLs), and chromosome segment substitution lines (CSSLs) from several different donors (reviewed by Mizobuchi et al., 2016). One major QTL, *qBSR11-kc*, showing 23.0%–25.9% of the total phenotypic variation was identified from *indica* variety CH45 (Matsumoto et al., 2017).

4.6 Narrow brown leaf spot also called "narrow brown spot" (pathogen: *S. oryzina*, syn. *C. janseana*, *C. oryzae*)

The genetic architecture of narrow brown spot (narrow brown leaf spot) resistance was almost unknown. A recent genetic analysis using the RIL population derived from the cross between two US varieties (Cypress and LaGrue) identified a single large-effect QTL, *CRSP-2.1*, explaining 81.4% of the phenotypic variation (Addison et al., 2021). The causal gene is not confirmed yet, but the major QTL might be useful in a breeding program.

4.7 Aggregate sheath spot (pathogen: *R. oryzae-sativae*)

Aggregate sheath spot disease has been reported in many Asian countries, as well as the USA, South America, and Australia, and it can cause ~20% of yield loss (Lanoiselet et al., 2007). Good levels of resistance to aggregate sheath spot were identified from *O. rufipogon* and successfully transferred into cultivars (McKenzie et al., 1994). Recent GWAS with tropical *japonica* and *indica* populations identified a handful of QTLs (Rosas et al., 2018).

4.8 Sheath rot (pathogen: *S. oryzae*)

Rice sheath rot diseases are found in most rice-growing areas of the world and cause 20%–85% ranges of yield losses, making it an emerging ubiquitous destructive disease of rice (Bigirimana et al., 2015). However, rice sheath rot is less studied, and no reliable germplasm or genetic factors have been identified yet.

4.9 Stem rot (pathogen: *S. oryzae*, syn. *N. oryzae*)

Stem rot disease resistance was found in wild rice species (*O. nivara* and *O. rufipogon*) and weedy rice (*O. sativa* f. *spontanea*) (Figoni et al., 1983), and the stem rot resistance was successfully transferred from *O. rufipogon* to California rice cultivars through interspecific hybridization (Oster, 1992). Recently, several QTLs for stem rot resistance were identified from *indica* germplasm through a GWAS analysis (Rosas et al., 2018).

5 Bacterial diseases and available genetic resources

Rice productions are significantly affected by several major bacterial diseases: BB (pathogen: *Xanthomonas oryzae* pv. *oryzae*

(*Xoo*)), "bacterial leaf streak" (BLS) (pathogen: *X. oryzae* pv. *oryzicola* (*Xoc*)), "bacterial sheath brown rot" also called "rice sheath rot" (pathogen: *Pseudomonas fuscovaginae*), and "bacterial seedling rot" (BSR), and "bacterial grain rot" (BGR) caused by the same pathogen (*Burkholderia glumae*). To date, a handful of genes have been cloned for BB resistance, but none yet for other bacterial diseases. Here, we described BB resistance genes cloned and some genetic resources for other bacterial pathogens.

5.1 Bacterial blight (pathogen: *X. oryzae* pv. *oryzae* (*Xoo*))

Among the bacterial diseases, BB caused by *Xoo* is the most destructive bacterial disease in rice. Thus, it has been intensively studied for the isolation of BB-resistant germplasm, genetic analysis, gene identification, and molecular mechanism of wars between *Xoo* and rice. To date, at least 47 *Xoo* resistance QTLs and genes (named *Xa* genes) have been identified from diverse germplasms, including cultivated rice, rice mutant lines, and wild rice species. *Xa21* from *Oryza longistaminata* introgression line (IRBB21) was first cloned in 1995 by Song et al. and followed by *Xa1* from the IRBB1 line (Yoshimura et al., 1998). Later, *Xa2*, *Xa31(t)*, *CGS-Xo1*, *Xa14*, and *Xa45(t)* were identified as a group of *Xa1* allelic R genes (Ji et al., 2020). Currently, 13 different genes/loci consisting of ~23 allelotypes have been cloned and characterized (Table 3), that is, *Xa1*/*Xa2*=*Xa31(t)*/*Xa14*/*Xa45(t)*/*CGS-Xo1*, *Xa3*=*Xa26*, *Xa4*, *xa5*, *Xa7*, *Xa10*, *xa13*/*OsSWEET11*/*Os8N3*, *Xa21*, *Xa23*, *Xa47(t)*, *xa25*/*OsSWEET13*/*OsMtN3*, *Xa27*, and *xa41(t)*/*OsSWEET14*/*Os11N3*. The 13 cloned BB resistance genes encode several types of proteins: NBS-LRR (*Xa1*/*Xa1* alleles and *Xa47(t)*), leucine-rich repeat receptor-like kinases (LRR-RLKs) (*Xa3*=*Xa26* and *Xa21*), a cell wall-associated kinase (WAK) (*Xa4*), executor R proteins (*Xa7*, *Xa10*, *Xa23*, and *Xa27*), SWEET/sugar transporter proteins (*xa13*/*OsSWEET11*, *xa25*/*OsSWEET13*, and *xa41(t)*/*OsSWEET14*), and a transcription factor gamma subunit protein (*xa5*). The genes encoding NBS-LRR, LRR-RLK, and WAK are involved in pathogen recognition and activation of the innate immune system, whereas the genes encoding executor R proteins are transcriptionally activated by the *Xoo* transcription activator-like (TAL) effector protein and trigger programmed cell death (PCD)-based hypersensitive response (HR). Thus, for the genes mentioned above, the functional alleles from the BB-resistant donor sources are dominant. In contrast, BB resistance is caused by sequence mutations at the TAL effector binding sites in the promoter of the SWEET (Sugar Will Eventually be Exported

TABLE 3 The cloned bacterial blight resistance genes.

Gene	Chr	Location (bp)	MSU_ID	RAPDB_ID	Encoding protein	Resistant/donor allele	Inheritance pattern of R-allele	Reference
Xa1/*Xa2*=*Xa31(t)*/*Xa14*/*Xa45(t)*/*CGS-Xo1*	4	31,638,099	Os04g53120	Os04g0622600	NBS-LRR	IRBB1 (*Xa1*), IRBB2 (*Xa2*), IRBB14 (*Xa14*), Zhachanglong (*Xa31(t)*), Carolina Gold Select (*CGS-Xo1*), *Oryza nivara* IRGC102463 (*Xa45(t)*)	Dominant	Yoshimura et al., 1998; Ji et al., 2020
xa5	5	437,043	Os05g01710	Os05g0107700	Transcription factor IIA gamma subunit	IRBB5	Recessive	Blair et al., 2003; Iyer and McCouch, 2004
Xa27	6	23,653,851	Os06g39810	Os06g0599600	Executor R protein	IRBB27/*Oryza minuta* IRGC101141	Dominant	Gu et al., 2005
Xa7$^{(a)}$	6	28,015,259	–	–	Executor R protein	IRBB7	Dominant	Chen et al., 2021; Wang et al., 2021
xa13/*OsSWEET11*/*Os8N3*	8	26,725,952	Os08g42350	Os08g0535200	SWEET-type protein	IRBB13	Recessive	Chu et al., 2006
xa41(t)/*OsSWEET14*/*Os11N3*	11	18,171,707	Os11g31190	Os11g0508600	SWEET-type protein	African wild and cultivated rice species *Oryza barthii* and *Oryza glaberrima*	Recessive	Hutin et al., 2015
Xa21	11	21,277,443	Os11g36180	Os11g0569733	LRR receptor kinase-like protein	IRBB21 (*Oryza longistaminata* IL)	Dominant	Song et al., 1995
Xa10	11	22,181,556	Os11g37570	Os11g0586400	Executor R protein	IRBB10, CAS209	Dominant	Tian et al., 2014
Xa23	11	22,204,131	–	Os11g0586701	Executor R protein	CBB23/*Oryza rufipogon*	Dominant	Wang C, et al., 2015

(Continued)

TABLE 3 Continued

Gene	Chr	Location (bp)	MSU_ID	RAPDB_ID	Encoding protein	Resistant/donor allele	Inheritance pattern of R-allele	Reference
Xa47(t)	11	27,983,597	Os11g46200	Os11g0688832	NBS-LRR	*O. rufipogon*	Dominant	Xing et al., 2021
Xa4[(b)]	11	28,357,055	Os11g47140	Os11g0694100	cell wall-associated kinase (WAK)	IRBB4	Dominant	Hu et al., 2017
Xa3=Xa26	11	28,399,360	Os11g47210	–	LRR receptor kinase-like protein	Minghui 63, IRBB3	Dominant	Sun et al., 2004
xa25/ OsSWEET13/ OsMtN3	12	17,302,127	Os12g29220	Os12g0476200	SWEET-type protein	Minghui 63	Recessive	Liu et al., 2011

[(a)]The sequence of Xa7 is completely absent in the Nipponbare reference genome (IRGSP1.0) and also most of japonica varieties. Thus, the location of the closest marker (M10) to Xa7 by Chen et al. (2021) is given in the above table.
[(b)] The sequence of xa4 gene was not fully aligned in the Nipponbare reference genome (IRGSP1.0). Thus, the information of the highest homology sequence was described above.

Transporter) genes and thus a recessive allele. BB resistance of *xa5* gene relies on one amino acid difference between resistance and susceptible lines in *Xa5* protein (a general eukaryotic transcription factor), and the BB-resistant allele is recessive (Iyer and McCouch, 2004). The cloned 13 genes are distributed on six chromosomes (one gene each on Chr 4, 5, 8, and 12; two genes on Chr 6; six genes on Chr 11) (Table 3, Figure 1). Six cloned genes on Chr 11 are closely located to each other in ~10.2-Mb size (18.2–28.4-Mb region on Chr 11) (Figure 1). Thus, in the case of gene pyramiding using the six genes on Chr 11, breeders need to consider producing enough progenies for obtaining pyramided alleles that occur by recombination between two closely located genes. Several cloned genes, including *Xa7*, *Xa23*, *xa41*, and *Xa47*, were reported as broad-spectrum resistance genes/alleles (Liu et al., 2021). NILs with single BB resistance genes were developed through IRRI-Japan collaboration designated as "IRBB" lines (Ogawa et al., 1991), and additional NILs (IRBB) with single or multiple BB resistance genes (two to five genes) were developed in the BB-susceptible background IR24 at IRRI, Philippines. Differential reactions of the NILs (IRBB lines) with single and pyramided *Xa* genes to 11 races in the Philippines were observed, and the results are available at the IRRI Rice knowledge bank (http://www.knowledgebank.irri.org/ricebreedingcourse/Breeding_for_disease_resistance_Blight.htm). The IRBB lines possessing multiple *Xa* genes (two to five genes) exhibited broad-spectrum resistance than the single gene introgression IRBB lines. Similarly, pyramiding of *Xa* genes such as *Xa21* + *Xa33*, *Xa21* + *xa13* + *xa5*, and *Xa4* + *xa5* + *Xa7* + *xa13* + *Xa21* offers greater and broader resistance to *Xoo* than an individual resistance gene (Pradhan et al., 2015; BalachIranjeevi et al., 2018; Hsu et al., 2020). The IRBB sets were also tested with 16 isolates in Korea, and the results showed that *xa5* was strong and broad-spectrum

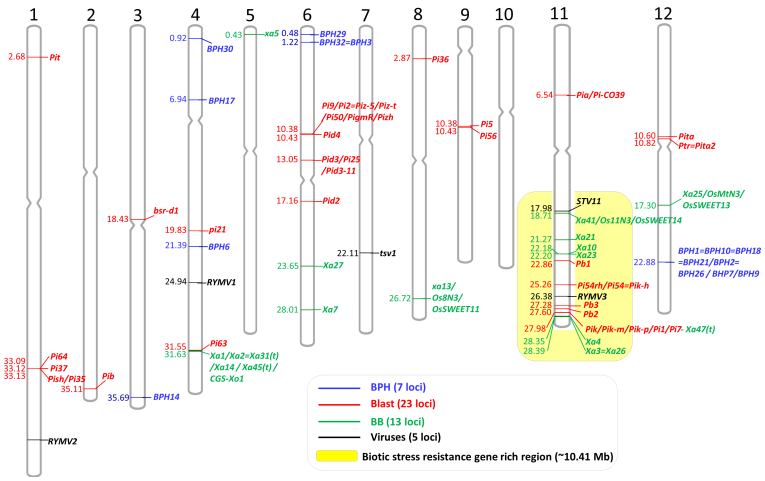


FIGURE 1
Physical locations of the 48 cloned genes conferring biotic stress resistance in rice. The cloned genes were mapped on the rice reference genome (Os-Nipponbare-Reference-IRGSP-1.0). Blue, red, green, and black bars mean brown planthopper (BPH), blast, bacterial blight, and virus resistance genes, respectively. Biotic stress resistance gene-rich region was highlighted by yellow background (out of 48 genes, 14 genes were on the 10.41-Mb region of the long arm of Chr 11).

resistant than any other *Xa* genes (Jeung et al., 2006). Rice possessing *Xa7* exhibited less disease than lines without *Xa7* over 11 years in the Philippines, even though the virulence of *Xoo* field populations increased. In addition, *Xa7* restricted disease more effectively at high temperatures, while other *Xa* genes were less effective at high temperatures (Webb et al., 2010). The IRBB lines and stacked information including gene reactions, spectrum, durability, and influence of environments will be useful to select suitable genes/alleles for regional/local breeding programs and also for the development of durable and broad-spectrum resistant rice varieties.

5.2 Bacterial leaf streak (pathogen: *X. oryzae* pv. *oryzicola* (*Xoc*))

For BLS resistance, a handful of QTLs with low-to-moderate PVEs (2.64%–15.93%) were identified (Tang et al., 2000). In addition, a recent GWAS using 510 diverse rice accessions identified 79 quantitative trait nucleotides (QTNs) reflecting 69 QTLs for BLS resistance (Xie et al., 2021). However, no BLS resistance gene has been cloned yet. Among the BLS-resistant QTLs, the highest effect QTL, *qBlsr5a* (12.84%–15.93% PVE), was fine-mapped to 30.0-kb interval on Chr 5, and the resistant parent allele of *Os05g01710* gene within the fine-mapped region was identical to *xa5*, which is one of major BB resistance genes, suggesting that *Os05g01710* (*xa5*) is possibly the candidate gene of *qBlsr5a* (Xie et al., 2014).

5.3 Bacterial sheath brown rot also called rice sheath rot (pathogen: *P. fuscovaginae*)

"Rice sheath rot" disease symptoms can be caused by the bacterial pathogen "*P. fuscovaginae*" and also by the fungal pathogen "*S. oryzae*". A recent pathobiomes study revealed that *P. fuscovaginae* and *S. oryzae* were prevalent in symptomatic rice samples in highland and lowland, respectively, in Burundi, indicating that the pathogens exist independently and are not part of a complex disease (Musonerimana et al., 2020). However, no reliable resistant germplasm and genetic factors have been identified yet.

5.4 Bacterial panicle blight, bacterial seedling rot, and bacterial grain rot (pathogen: *B. glumae*)

Bacterial panicle blight (BPB), BSR, and BGR are caused by the same bacterial pathogen, *B. glumae*. It was first reported as BGR in Japan in 1955. Since then, BPB has been found in more than 18 countries globally including Asia, Africa, and North and South America (Zhou, 2019; Ortega and Rojas, 2021). Although it is an emerging disease globally, only several cultivars with partial resistance and 12 QTLs associated with partial resistance have been reported (Zhou, 2019). Regarding BSR resistance, one QTL (*RBG1*/*qRBS1*) was identified from the CSSL population (Nona Bokra introgressions in Koshihikari background) (Mizobuchi et al., 2016). For BGR resistance, 13 QTLs have been found from the two mapping populations: a BIL from Kele (R) × Hitomebore (S) and a RIL from TeQing (R) × Lemont (S) (Mizobuchi et al., 2016).

6 Viral diseases and available genetic resources

Seventeen rice viruses have been reported, including rice black-streaked dwarf virus (RBSDV), rice bunchy stunt virus (RBSV), rice dwarf virus (RDV), rice gall dwarf virus (RGDV), rice giallume virus (RGV), RGSV, rice hoja blanca virus (RHBV), rice necrosis mosaic virus (RNMV), RRSV, rice stripe necrosis virus (RSNV), rice stripe virus (RSV), rice transitory yellowing virus (RTYV) also named as rice yellow stunt virus (RYSV), rice tungro bacilliform virus (RTBV), rice tungro spherical virus (RTSV), RYMV, southern rice black-streaked dwarf virus (SRBSDV), and rice stripe mosaic virus (RSMV) (Hibino, 1996; Qin et al., 2019). Since most of the above viruses are arthropod-borne, damages may become more severe as the population of vector insects increases. Among the rice virus diseases, rice tungro disease (RTSV and RTBV), RYMV, and RSV have been causing serious yield loss in South/Southeast Asia, Africa, and temperate Asia, respectively. Thus, a few genes providing resistance to the major viruses above have been cloned. The use of viral disease resistance may significantly reduce the damage of viral diseases. In addition to this, the management of corresponding vector insects may mitigate the damage of viral diseases in the field.

6.1 Rice tungro disease caused by RTSV and RTBV

Rice tungro disease is a serious threat to rice production in South and Southeast Asia. Tungro disease viruses are transmitted from tungro-infected plant to another by leafhoppers. The most efficient vector is the green leafhopper (IRRI Rice Knowledge Bank). Tungro was found to be associated with two distinct viruses: RTSV and RTBV. A series of large-scale outbreaks of tungro were recorded in India, Thailand, Indonesia, Malaysia, the Philippines, Thailand, China, and Bangladesh. Tungro, as one of the destructive diseases of rice, causes yield losses of 5% to 10% annually and is estimated to cause an annual loss in rice production of approximately 1.5 billion US dollars worldwide (Dai and Beachy, 2009). In the late 1990s, several tungro-resistant sources, including landrace and wild species, were isolated and used in the breeding program by IRRI, and the most promising breeding lines were developed by crossing with Utri Merah donor (Azzam and Chancellor, 2002). Afterward, Encabo et al. (2009) revealed that RTBV and RTSV are inherited separately from rice accession Utri Merah, conferring resistance to both RTBV and RTSV, and Lee et al. (2010) cloned the causal recessive gene (named as *tsv1*) involved in RTSV resistance in Utri Merah. *TSV1* encodes eukaryotic translation initiation factor 4G (eIF4G), and mutation

on the protein-coding sequence of *TSV1* in Utri Merah (*tsv1* allele) may impair the RTSV RNA translation, resulting in tungro resistance. The *tsv1*-Utri Merah allele is widely used for tungro resistance improvement in many breeding programs.

6.2 Rice yellow mottle virus

Since RYMV was first discovered in Kenya in 1970, it has been reported from only the countries in the African continent. RYMV causes the most serious damage in Africa among all the rice diseases. Primary infection of RYMV in rice fields is mediated by beetle family chrysomelids, and secondary spread occurs mainly through mechanical contact between infected and healthy leaves by wind (Kouassi et al., 2005). In the past, farmers have been advised to use chemicals to eliminate beetle vectors. The most effective and sustainable way to manage RYMV is to use tolerant and resistant varieties (Abo et al., 1997).

High RYMV resistance was found in one African rice cultivar (*Oryza glaberrima*), Tog5681, and one *O. sativa* cultivar, Gigante. Evaluation of the crosses of these two highly RYMV-resistant cultivars suggests the presence of a single recessive gene (Ndjiondjop et al., 1999). Later, it was discovered that the gene is *RYMV1*, and the gene encodes a eukaryotic translation initiation factor, eIF4(iso)4G (Albar et al., 2006). In sequence comparisons with the dominant susceptible allele (*Rymv1-1*), four different recessive resistant alleles from one *O. sativa* var. Gigante (*rymv1-2*) and three *O. glaberrima* accessions (Tog5681 (*rymv1-3*), Tog5672 (*rymv1-4*), and Tog5674 (*rymv1-5*)) were characterized by the presence of short amino acid substitutions or short deletions in the MIF4G domain of the protein (Albar et al., 2006; Thiémélé et al., 2010). Allele-specific markers targeting mutations or deletions characterizing different *RYMV1* were developed for improving MAS for the introduction of the resistance alleles into susceptible cultivars of *O. sativa* or *O. glaberrima* (Thiémélé et al., 2010). In the second major recessive resistance gene, *RYMV2*, it was identified that 1-bp deletion on the coding sequence of the rice homolog of the *Arabidopsis CPR5* gene, known to be a defense mechanism regulator, from the resistant African rice (*O. glaberrima*) Tog7291 provided RYMV resistance (Orjuela et al., 2013). A single dominant resistant gene *RYMV3* encoding NBS-LRR protein was identified from the *O. glaberrima* Tog5307 (Pidon et al., 2017). Novel resistant alleles and accessions for *RYMV2* and *RYMV3* were identified by screening 268 *O. glaberrima* accessions and sequencing (Pidon et al., 2020), and five new resistant germplasm were isolated from Korean rice lines (Asante et al., 2020). The cloned genes with different resistant alleles will be useful to improve RYMV resistance, especially for the breeding program for the African continent.

6.3 Rice stripe virus

RSV is an RNA-type virus belonging to the genus *Tenuivirus*, and it is transmitted by SBPHs. RSV has been reported only in China, Japan, Korea, and Taiwan, where *japonica* rice is cultivated, and it caused severe damage to the rice fields in Eastern China, Japan, and Korea. While most *indica* varieties are resistant to RSV, the majority of *japonica* varieties are highly susceptible. A number of RSV-resistant QTLs have been reported from diverse *indica*-resistant donors, and the major QTLs were repeatedly detected on Chr 11 among several QTL mapping (Cho et al., 2013). Finally, the major QTL, *qSTV11*, originated from an *indica* variety Kasalath and was cloned (Wang et al., 2014). *STV11*-Kasalath allele encodes a sulfotransferase (OsSOT1) protein catalyzing the conversion of salicylic acid (SA) into sulfonated SA (SSA), whereas the protein encoded by the susceptible allele *STV11* loses this activity. *STV11* gene will be useful in improving RSV resistance in the *japonica* varieties.

7 Physical locations of the cloned genes/alleles on rice chromosomes

Graphical mapping of the cloned genes on 12 rice chromosomes will be useful information for MAS breeding, especially for gene pyramiding, as well as mapping new biotic stress resistance genes. We mapped the physical locations of all the cloned 48 biotic stress resistance genes on the 12 rice chromosomes (Figure 1). The cloned genes were not evenly distributed across the rice genome. No biotic stress resistance gene was cloned yet on Chr 10. In contrast, Chr 11 possesses the highest number of genes (15 genes), following Chr 6 (eight genes), Chr 4 (seven genes), and Chr 12 (four genes), with these four chromosomes harboring 34 genes out of 48 cloned genes (70.83%). Interestingly, 14 cloned genes associated with blast, bacterial blight, and virus resistance were on the 10.41-Mb region of the long arm of Chr 11 (Chr 11: 17.98–28.39 Mb), and it took 29.16% of the cloned genes. Biotic stress resistance genes are ~10 times more enriched in this specific region than any other loci (the expected distribution is ~1.2 cloned gene/10 Mb). Another interesting point is that the bacterial blight resistance gene *Xa47 (t)* (*Os11g46200*) encoding NBS-LRR is overlapped with the blast resistance gene *Pik*/*Pik-m*/*Pik-p*/*Pi1*/*Pi7* consisting of two NBS-LRR genes (*Os11g46200* and *Os11g46210*). In some loci, different resistance alleles at the same locus, such as *BPH1* locus, *Pi9* locus, *Pik* locus, and *Xa1* locus, were identified (Tables 1–3). Although some of them among the alleles showed different reactions to pathotypes, unfortunately, they cannot be pyramided by MAS due to the same physical location among the alleles. Thus, breeders need to choose one suitable allele among the alleles based on the regional pathotypes/isolates. Similarly, in gene pyramiding/stacking, breeders should also consider the physical distance between/among the target genes. If the two target genes are closely located with each other (<~1Mb) on the same chromosome (for example, *Xa10* and *Pb1* on Chr 11, *Pik* and *Xa4* on Chr 11, and *Pita* and *Ptr=Pita2* on Chr 12; see Figure 1), breeders need to produce many progenies to obtain the gene pyramided plants through the selection of the recombinant plants between the two target loci. In rice, a handful of recombination hot and cold regions are reported, and the average recombination frequency is approximately 4.35 cM per Mb (Si et al., 2015). In addition, breeders also need to check the target loci whether the important genes governing other agronomic traits are present near the target biotic stress resistance gene to avoid

linkage drag. For instance, a key amylose synthesis gene *Waxy*/*GBSS1* (1.76-Mb location on Chr 6) is tightly linked with *BPH32* (1.22 Mb on Chr 6), and a major heading date gene *Hd1* (9.33 Mb on Chr 6) is closely located with *Pi2* gene (10.38 Mb on Chr 6). Thus, breeders should consider the locations of the important agronomic traits genes near the target genes, especially when the breeders try to retain the original characteristics of the elite background variety, except for the target biotic stress resistance. A map of the physical locations of the cloned genes (Figure 1) will be helpful for consideration of the above points in MAS breeding programs.

8 Available DNA markers for MAS breeding

DNA markers are essential tools for genetic analysis as well as marker-assisted breeding. We tried to collect all the markers published and used in the previous breeding programs, and we collected ~500 markers in total for the cloned biotic stress resistance genes (Table S1). We filed essential information on the markers for the potential users, including marker types (InDel, CAPS/dCAPS, dominant markers, and tetra-primer method markers) and primer sequences. Also, we cited the original references of each marker so that breeders can obtain detailed and additional information if needed. Furthermore, we mapped the location of all the markers in the rice reference genome sequence (Os-Nipponbare-Reference-IRGSP-1.0: https://rapdb.dna.affrc.go.jp/). This information provides physical distance between the target gene and the markers, and it will be helpful to reduce the selection of false positives during MAS. For examples, some markers for the *BPH1*, *BP17*, *xa13*, *Xa27*, *Pi9*, *Piz-t*, *Pizh*, *Pish*, *Pi5*, *Pita2*, and *RYMV1* genes/alleles are a bit far (>1 Mb) from the gene locus (Table S1). Selection of genic or gene-tightly linked markers would reduce false-positive selection. In cases of multi-alleles for the same gene, such as *BPH1* and *Pi9*, all the available markers for the same gene can be tested to check the possibility of polymorphism between the parents, and the selected polymorphic markers can be used in MAS breeding (for example, *BPH18* markers for *BPH26* MAS breeding). All the information on the markers is described in Table S1.

9 Conclusions and perspective

In this review, we summarized all the cloned genes associated with biotic stress resistance (Tables 1–4), mapped the physical location of the genes on 12 rice chromosomes (Figure 1), and consolidated the available markers associated with the cloned genes (Table S1). Furthermore, we also briefly introduced genetic resources such as QTLs and donor sources for some biotic stress if the cloned genes are not available yet. The information presented in this review will be helpful for checking the available genetic resources for biotic stress resistance and also for MAS breeding for the genetic improvement of biotic stress resistance in rice. As shown in many previous reports, pyramiding of QTLs/genes might be a practical solution to breed durable and broad-spectrum resistant varieties.

Approximately 48 genes, which are natural alleles and provide biotic stress resistance, have been cloned only for the major biotic stresses, including BPH, blast, BB, and some viruses. However, no genes have been cloned yet for other biotic stresses. Preparation of the reliable genetic factors (genes/QTLs) associated with currently problematic and emerging pathogens is very important for stable high-yield rice production, and thus, scientists/geneticists need to put much effort into this pending issue. Screening wild relatives of rice in the genus *Oryza* will be one of the ideal approaches. Many biotic stress resistance genes were already cloned from wild germplasm (see Tables 1–3), such as *BPH14* (*O. officinalis*), *Pi9* (*Oryza minuta*), and *Xa21* (*O. longistaminata*). More than 4,500 accessions of wild rice species are stored in the IRRI Genebank (Banaticla-Hilario and Sajise, 2022), and most of the germplasms were not screened yet. Recently, a genome-wide InDel marker set (475 polymorphic markers) discriminating the alleles between *O. sativa* and the other seven AA-genome *Oryza* species was developed

TABLE 4 The cloned virus resistance genes.

Gene	Chr	Location (bp)	MSU_ID	RAPDB_ID	Encoding protein	Resistant/donor allele	Inheritance pattern of R-allele	Reference
tsv1	7	22,114,961	Os07g36940	Os07g0555200	Eukaryotic translation initiation factor 4G (eIF4G)	Utri Merah (UM82)	Recessive	Lee et al., 2010
RYMV1	4	24,946,171	Os04g42140	Os04g0499300	Eukaryotic translation initiation factor isoform 4G-1 (eIF(iso)4G1)	*Oryza sativa* Gigante (*rymv1-2*)/*Oryza glaberrima* accessions Tog5681, Tog5672, and Tog5674 for *rymv1-3*, *rymv-4*, and *rymv-5*, respectively	Recessive	Albar et al., 2006; Thiémélé et al., 2010
RYMV2	1	40,073,727	Os01g68970	Os01g0918500	Constitutive expresser of PR genes5 (CPR5)	*O. glaberrima* Tog7291	Recessive	Orjuela et al., 2013
RYMV3	11	26,380,866	Os11g43700	Os11g0657900	NBS-LRR	*O. glaberrima* Tog5307	Dominant	Pidon et al., 2017
STV11	11	17,985,011	Os11g30910	Os11g0505300	Sulfotransferase (OsSOT1)	Kasalath	ND	Wang et al., 2014

ND, not determined.

to harness AA-genome wild species (Hechanova et al., 2021). The genes identified from wild germplasm will be rare alleles due to mostly untapped and unused materials in breeding, and thus, they will be effective in most *indica* and *japonica* backgrounds.

The incidence of pathogens and insect pests will change in time and space; notably, it will be also influenced by climate changes. As examples, some BPH resistance genes were affected by artificial climate change conditions (the atmospheric temperature with corresponding carbon dioxide at the ambient, year 2050 and year 2100) (Kuang et al., 2021) and also by nitrogen fertilizer treatments (Lin et al., 2022). Moreover, most of the genes/QTLs reported were tested with limited numbers of isolates/biotypes, which were collected in specific locations and years. Thus, the identified genes/QTLs could not guarantee resistance across locations, time, and environments. Testing donor germplasm, especially sets of NILs possessing specific genes/QTLs such as NILs for BPH (Jena et al., 2017), blast (Telebanco-Yanoria et al., 2010; Telebanco-Yanoria et al., 2011; Fukuta et al., 2022), and BB (Ogawa et al., 1991; IRBB lines), with prevalence races/biotypes in the target regions, would be a good strategy to select effective genes/alleles in breeding program.

DNA markers are essential tools for genetic analysis and breeding. DNA markers could reduce the time and effort in developing and improving biotic-resistant cultivars through marker-assisted breeding. Due to the accessibility and technical simplicity for the rice breeders, most of the markers are PCR and gel-based markers, including SSR (RM) markers, InDel markers, CAPS markers, tetra-primer PCR markers, and dominant PCR markers (Table S1). These markers have contributed much to MAS breeding. However, the gene/allele-specific markers are limited to some specific genes, and a high portion of the markers are the gene-linked makers (sometimes more than a few Mb distance from the gene), probably causing that false-positive selection in MAS breeding. Thus, breeders should check the marker–gene linkage (distance between the gene and markers) and also marker quality (reproducibility and polymorphism between parents) before starting MAS breeding. For efficient and precious introgression of the target genes, currently, available markers might be insufficient. Developments of breeder-friendly allele-specific markers and enough number of polymorphic markers with high reproducibility for many biotic stress resistance genes/alleles are urgently needed. This will help the rapid deployment of target biotic stress resistance genes in the elite local varieties.

In addition to MAS breeding, CRISPR-based genome editing technologies might be an alternative solution for the fast improvement of biotic stress resistance. The advantage of genome editing is that the techniques can directly improve target traits in elite backgrounds without crossing with the donor lines. Thus, some unexpected phenotypes caused by linkage drag or other donor introgressions happening during MAS breeding will not be considered in genome editing-based trait improvement. Recently, its potential was already shown in BB resistance improvement by CRISPR-based promoter editing of three *SWEET* genes in rice (Oliva et al., 2019) and in tungro virus resistance by editing of *TSV1* gene (Macovei et al., 2018). Another advantage is that genome-edited products are regulated with lesser stringency in many countries compared to conventional genetically modified organisms (GMOs). Together with cross-based breeding, genome editing technologies can contribute fast genetic improvement of target traits in the elite variety backgrounds without linkage drag and other donor introgressions.

Author contributions

JH, C-PL, AT, E-KA, JJ, I-RC, RS, KJ, and S-RK conceived this review paper. ES, SH, I-RC, and S-RK performed the literature search and wrote the draft. The manuscript was improved by revisions by all the authors. All authors agreed to the published version of the manuscript.

Funding

The preparation and publication of this review paper were supported by the Temperate Rice Research Consortium (TRRC) project and the bilateral projects for biotic stress resistance improvement in rice between the Gene Identification and Validation (GIV) group of the International Rice Research Institute (IRRI), Philippines, and the national agricultural research and extension systems (NARES) including Taiwan Agricultural Research Institute (TARI, Taiwan), General Directorate of Agricultural Research and Policies (GDAR, Turkey), Rural Development Administration (RDA, Korea), Rice Department (RD, Thailand), and Indian Council of Agricultural Research (ICAR, India).

Acknowledgments

We are thankful to Dr. Van Schepler Luu and Dr. Gilda Jonson from the IRRI pathology group for the careful editing of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1247014/full#supplementary-material

References

Abo, M. E., Sy, A. A., and Alegbejo, M. D. (1997). Rice yellow mottle virus (RYMV) in Africa: evolution, distribution, economic significance on sustainable rice production and management strategies. *J. Sustain. Agric.* 11 (2-3), 85–111. doi: 10.1300/J064v11n02_08

Addison, C. K., Angira, B., Cerioli, T., Groth, D. E., Richards, J. K., Linscombe, S. D., et al. (2021). Identification and mapping of a novel resistance gene to the rice pathogen, *Cercospora janseana*. *Theor. Appl. Genet.* 134, 2221–2234. doi: 10.1007/s00122-021-03821-2

Albar, L., Bangratz-Reyser, M., Hébrard, E., Ndjiondjop, M. N., Jones, M., and Ghesquière, A. (2006). Mutations in the eIF (iso) 4G translation initiation factor confer high resistance of rice to rice yellow mottle virus. *Plant J.* 47 (3), 417–426. doi: 10.1111/j.1365-313X.2006.02792.x

Andargie, M., Li, L., Feng, A., Zhu, X., and Li, J. (2018). Mapping of the quantitative trait locus (QTL) conferring resistance to rice false smut disease. *Curr. Plant Biol.* 15, 38–43. doi: 10.1016/j.cpb.2018.11.003

Asante, M. D., Amadu, B., Traore, V. S. E., Oppong, A., Adebayo, M. A., Aculey, P., et al. (2020). Assessment of Korean rice lines for their reaction to rice yellow mottle virus in Ghana. *Heliyon* 6 (11), e05551. doi: 10.1016/j.heliyon.2020.e05551

Ashikawa, I., Hayashi, N., Yamane, H., Kanamori, H., Wu, J., Matsumoto, T., et al. (2008). Two adjacent nucleotide-binding site–leucine-rich repeat class genes are required to confer Pikm-specific rice blast resistance. *Genetics* 180 (4), 2267–2276. doi: 10.1534/genetics.108.095034

Ashkani, S., Rafii, M. Y., Shabanimofrad, M., Ghasemzadeh, A., Ravanfar, S. A., and Latif, M. A. (2016). Molecular progress on the mapping and cloning of functional genes for blast disease in rice (*Oryza sativa* L.): current status and future considerations. *Crit. Rev. Biotechnol.* 36 (2), 353–367. doi: 10.3109/07388551.2014.961403

Azzam, O., and Chancellor, T. C. (2002). The biology, epidemiology, and management of rice tungro disease in Asia. *Plant Dis.* 86 (2), 88–100. doi: 10.1094/PDIS.2002.86.2.88

BalachIranjeevi, C., Naik, B., Kumar, A., Harika, G., Hajira, S., and Kumar, D. (2018). Marker-assisted pyramiding of two major broad-spectrum bacterial blight resistance genes, *Xa21* and *Xa33* into an elite maintainer line of rice, DRR17B. *PLoS One* 13 (10), e0201271. doi: 10.1371/journal.pone.0201271

Banaticla-Hilario, M. C. N., and Sajise, A. G. (2022). "Recent developments in wild rice conservation, research, and use". in *Plant Genet. Resources Inventory Collection Conserv.* eds. S. Ramamoorthy, I.J. Buot and R. Chandrasekaran (Singapore: Springer), 43–76. doi: 10.1007/978-981-16-7699-4_3

Bandumula, N. (2018). Rice production in Asia: key to global food security. *Proc.Natl. Acad. Sci.* 88, 1323–1328. doi: 10.1007/s40011-017-0867-7

Bebber, D. P. (2015). Range-expanding pests and pathogens in a warming world. *Annu. Rev. Phytopathol.* 53, 335–356. doi: 10.1146/annurev-phyto-080614-120207

Bentur, J. S., Rawat, N., Divya, D., Sinha, D. K., Agarrwal, R., Atray, I., et al. (2016). Rice–gall midge interactions: battle for survival. *J. Insect Physiol.* 84, 40–49. doi: 10.1016/j.jinsphys.2015.09.008

Beser, N., Del Valle, M. M., Kim, S. M., Vinarao, B. R., Surek, H., and Jena, K. K. (2016). Marker-assisted introgression of a broad-spectrum resistance gene, Pi40 improved blast resistance of two elite rice (*Oryza sativa* L.). *Mol. Plant Breed.* 7, 1–15. doi: 10.5376/mpb.2016.07.0033

Bigirimana, V. D. P., Hua, G. K., Nyamangyoku, O. I., and Höfte, M. (2015). Rice sheath rot: an emerging ubiquitous destructive disease complex. *Front. Plant Sci.* 6 1066. doi: 10.3389/fpls.2015.01066

Blair, M. W., Garris, A. J., Iyer, A. S., Chapman, B., Kresovich, S., and McCouch, S. R. (2003). High resolution genetic mapping and candidate gene identification at the *xa5* locus for bacterial blight resistance in rice (*Oryza sativa* L.). *Theoret. Appl. Genet.* 107, 62–73. doi: 10.1007/s00122-003-1231-2

Brown, J. K. (2002). Yield penalties of disease resistance in crops. *Curr. Opin. Plant Biol.* 5 (4), 339–344. doi: 10.1016/S1369-5266(02)00270-4

Bryan, G. T., Wu, K. S., Farrall, L., Jia, Y., Hershey, H. P., McAdams, S. A., et al. (2000). A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. *Plant Cell* 12 (11), 2033–2045. doi: 10.2307/3871103

Cabauatan, P. Q., Cabunagan, R. C., and Choi, I. R. (2009). "Rice viruses transmitted by the brown planthopper *Nilaparvata lugens* Stål.", in *Planthoppers: New threats to sustainability Intensive Rice production Syst. Asia* eds. K.L. Heong and B. Hardy (Philippines: International Rice Research Institute), 357–368.

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., et al. (2013). The rice resistance protein pair RGA4/RGA5 recognizes the *Magnaporthe oryzae* effectors AVR-Pia and AVR1-CO39 by direct binding. *Plant Cell* 25 (4), 1463–1481. doi: 10.1105/tpc.112.107201

Chaloner, T. M., Gurr, S. J., and Bebber, D. P. (2021). Plant pathogen infection risk tracks global crop yields under climate change. *Nat. Clim. Change* 11 (8), 710–715. doi: 10.1038/s41558-021-01104-8

Chen, H., He, H., Zou, Y., Chen, W., Yu, R., Liu, X., et al. (2011). Development and application of a set of breeder-friendly SNP markers for genetic analyses and molecular breeding of rice (Oryza sativa L.). *Theor. Appl. Genet.* 123, 869–879. doi: 10.1007/s00122-011-1633-5

Chen, X., Liu, P., Mei, L., He, X., Chen, L., Liu, H., et al. (2021). *Xa7*, a new executor R gene that confers durable and broad-spectrum resistance to bacterial blight disease in rice. *Plant Com* 2 (3), 100143. doi: 10.1016/j.xplc.2021.100143

Chen, X., Shang, J., Chen, D., Lei, C., Zou, Y., Zhai, W., et al. (2006). AB-lectin receptor kinase gene conferring rice blast resistance. *Plant J.* 46 (5), 794–804. doi: 10.1111/j.1365-313X.2006.02739.x

Chen, J., Shi, Y., Liu, W., Chai, R., Fu, Y., Zhuang, J., et al. (2011). A Pid3 allele from rice cultivar Gumei2 confers resistance to Magnaporthe oryzae. *J. Genet. Genomics* 38 (5), 209–216. doi: 10.1016/j.jgg.2011.03.010

Chen, Z., Zhao, W., Zhu, X., Zou, C., Yin, J., Chern, M., et al. (2018). Identification and characterization of rice blast resistance gene *Pid4* by a combination of transcriptomic profiling and genome analysis. *J. Genet. Genomics* 45 (12), 663–672. doi: 10.1016/j.jgg.2018.10.007

Cho, W. K., Lian, S., Kim, S. M., Park, S. H., and Kim, K. H. (2013). Current insights into research on Rice stripe virus. *Plant Pathol. J.* 29 (3), 223. doi: 10.5423/PPJ.RW.10.2012.0158

Chu, Z., Fu, B., Yang, H., Xu, C., Li, Z., Sanchez, A., et al. (2006). Targeting *xa13*, a recessive gene for bacterial blight resistance in rice. *Theoret. Appl. Genet.* 112, 455–461. doi: 10.1007/s00122-005-0145-6

Dai, S., and Beachy, R. N. (2009). Genetic engineering of rice to resist rice tungro disease. *In Vitro Cell. Dev. Biol. - Plant* 45, 517–524. doi: 10.1007/s11627-009-9241-7

Das, A., Soubam, D., Singh, P. K., Thakur, S., Singh, N. K., and Sharma, T. R. (2012). A novel blast resistance gene, *Pi54rh* cloned from wild species of rice, *Oryza rhizomatis* confers broad spectrum resistance to Magnaporthe oryzae. *Funct. Integrat. Genom.* 12, 215–228. doi: 10.1007/s10142-012-0284-1

Deng, Y., Zhai, K., Xie, Z., Yang, D., Zhu, X., Liu, J., et al. (2017). Epigenetic regulation of antagonistic receptors confers rice blast resistance with yield balance. *Science* 355 (6328), 962–965. doi: 10.1126/science.aai8898

Divya, D., Himabindu, K., Nair, S., and Bentur, J. S. (2015). Cloning of a gene encoding LRR protein and its validation as candidate gall midge resistance gene, *Gm4*, in rice. *Euphytica* 203, 185–195. doi: 10.1007/s10681-014-1302-2

Divya, D., Sahu, N., Nair, S., and Bentur, J. S. (2018). Map-based cloning and validation of a gall midge resistance gene, *Gm8*, encoding a proline-rich protein in the rice variety Aganni. *Mol. Biol. Rep.* 45 (6), 2075–2086. doi: 10.1007/s11033-018-4364-8

Dixit, S., Singh, U. M., Singh, A. K., Alam, S., Venkateshwarlu, C., Nachimuthu, V. V., et al. (2020). Marker assisted forward breeding to combine multiple biotic-abiotic stress resistance/tolerance in rice. *Rice* 13, 1–15. doi: 10.1186/s12284-020-00391-7

Du, B., Chen, R., Guo, J., and He, G. (2020). Current understanding of the genomic, genetic, and molecular control of insect resistance in rice. *Mol. Breed.* 40, 1–25. doi: 10.1007/s11032-020-1103-3

Du, B., Zhang, W., Liu, B., Hu, J., Wei, Z., Shi, Z., et al. (2009). Identification and characterization of *Bph14*, a gene conferring resistance to brown planthopper in rice. *Proc. Natl. Acad. Sci.* 106 (52), 22163–22168. doi: 10.1073/pnas.0912139106

Dyck, V. A., and Thomas, B. (1979). "The brown planthopper problem," in *Brown planthopper: threat to rice production in Asia* (Los Baños, Philippines: International Rice Research Institute), 3–17.

Encabo, J. R., Cabauatan, P. Q., Cabunagan, R. C., Satoh, K., Lee, J. H., Kwak, D. Y., et al. (2009). Suppression of two tungro viruses in rice by separable traits originating from cultivar Utri Merah. *Mol. Plant Microbe Interact.* 22 (10), 1268–1281. doi: 10.1094/MPMI-22-10-1268

Fahad, S., Adnan, M., Noor, M., Arif, M., Alam, M., Khan, I. A., et al. (2019). "Major constraints for global rice production," in *Advances in rice research for abiotic stress tolerance*. Eds. M. Hasanuzzaman, M. Fujita, K. Nahar and J. K. Biswas (United Kingdom: Woodhead Publishing), 1–22.

Figoni, R. A., Rutger, J. N., and Webster, R. K. (1983). Evaluation of wild *Oryza* species for stem rot (*Sclerotium oryzae*) resistance. *Plant Dis.* 67 (9), 998–1000. doi: 10.1094/PD-67-998

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., et al. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484 (7393), 186–194. doi: 10.1038/nature10947

Fujita, D., Kohli, A., and Horgan, F. G. (2013). Rice resistance to planthoppers and leafhoppers. *Crit. Rev. Plant Sci.* 32, 162–191. doi: 10.1080/07352689.2012.735986

Fukuoka, S., Saka, N., Koga, H., Ono, K., Shimizu, T., Ebana, K., et al. (2009). Loss of function of a proline-containing protein confers durable disease resistance in rice. *Science* 325 (5943), 998–1001. doi: 10.1126/science.117555

Fukuoka, S., Yamamoto, S. I., Mizobuchi, R., Yamanouchi, U., Ono, K., Kitazawa, N., et al. (2014). Multiple functional polymorphisms in a single disease resistance gene in rice enhance durable resistance to blast. *Sci. Rep.* 4 (1), 1–7. doi: 10.1038/srep04550

Fukuta, Y., Koide, Y., Kobayashi, N., Kato, H., Saito, H., Telebanco-Yanoria, M. J., et al. (2022). Lines for blast resistance genes with genetic background of Indica Group rice as international differential variety set. *Plant Breed.* 141 (5), 609–620. doi: 10.1111/pbr.13040

Gan, L., Zhai, C., and Hua, L. (2010). *Rice blast resistance gene Pi7 and application thereof. CN patent Application no: CN102094027A* (South China Agricultural University, China).

Goad, D. M., Jia, Y., Gibbons, A., Liu, Y., Gealy, D., Caicedo, A. L., et al. (2020). Identification of novel QTL conferring sheath blight resistance in two weedy rice mapping populations. *Rice* 13, 1–10. doi: 10.1186/s12284-020-00381-9

Gu, K., Yang, B., Tian, D., Wu, L., Wang, D., Sreekala, C., et al. (2005). R gene expression induced by a type-III effector triggers disease resistance in rice. *Nature* 435 (7045), 1122–1125. doi: 10.1038/nature03630

Guo, J., Xu, C., Wu, D., Zhao, Y., Qiu, Y., Wang, X., et al. (2018). *Bph6* encodes an exocyst-localized protein and confers broad resistance to planthoppers in rice. *Nat. Genet.* 50 (2), 297–306. doi: 10.1038/s41588-018-0039-6

Han, Y., Li, D., Yang, J., Huang, F., Sheng, H., and Sun, W. (2020). Mapping quantitative trait loci for disease resistance to false smut of rice. *Phytopathol. Res.* 2, 1–11. doi: 10.1186/s42483-020-00059-6

Hayashi, N., Inoue, H., Kato, T., Funao, T., Shirota, M., Shimizu, T., et al. (2010). Durable panicle blast-resistance gene *Pb1* encodes an atypical CC-NBS-LRR protein and was generated by acquiring a promoter through local genome duplication. *Plant J.* 64 (3), 498–510. doi: 10.1111/j.1365-313X.2010.04348.x

Hayashi, K., Yasuda, N., Fujita, Y., Koizumi, S., and Yoshida, H. (2010). Identification of the blast resistance gene *Pit* in rice cultivars using functional markers. *Theoret. Appl. Genet.* 121, 357–1367. doi: 10.1007/s00122-010-1393-7

Hechanova, S. L., Bhattarai, K., Simon, E. V., Clave, G., Karunarathne, P., Ahn, E. K., et al. (2021). Development of a genome-wide InDel marker set for allele discrimination between rice (*Oryza sativa*) and the other seven AA-genome *Oryza* species. *Sci. Rep.* 11 (1), 8962. doi: 10.1038/s41598-021-88533-9

Hibino, H. (1996). Biology and epidemiology of rice viruses. *Annu. Rev. Phytopathol.* 34, 249–274. doi: 10.1146/annurev.phyto.34.1.249

Hiremath, S. S., Bhatia, D., Jain, J., Hunjan, M. S., Kaur, R., Zaidi, N. W., et al. (2021). Identification of potential donors and QTLs for resistance to false smut in a subset of rice diversity panel. *Eur. J. Plant Pathol.* 159, 461–470. doi: 10.1007/s10658-020-02172-w

Horgan, F. G., Almazan, M. L. P., Vu, Q., Ramal, A. F., Bernal, C. C., Yasui, H., et al. (2019). Unanticipated benefits and potential ecological costs associated with pyramiding leafhopper resistance loci in rice. *Crop Prot.* 115, 47–58. doi: 10.1016/j.cropro.2018.09.013

Horgan, F. G., Bernal, C. C., Vu, Q., Almazan, M. L. P., Ramal, A. F., Yasui, H., et al. (2018). Virulence adaptation in a rice leafhopper: Exposure to ineffective genes compromises pyramided resistance. *Crop Prot.* 113, 40–47. doi: 10.1016/j.cropro.2018.07.010

Hsu, Y. C., Chiu, C. H., Yap, R., Tseng, Y. C., and Wu, Y. P. (2020). Pyramiding bacterial blight resistance genes in Tainung82 for broad-spectrum resistance using marker-assisted selection. *Int. J. Mol. Sci.* 21 (4), 1281. doi: 10.3390/ijms21041281

Hu, K., Cao, J., Zhang, J., Xia, F., Ke, Y., Zhang, H., et al. (2017). Improvement of multiple agronomic traits by a disease resistance gene via cell wall reinforcement. *Nat. Plants* 3 (3), 1–9. doi: 10.1038/nplants.2017.9

Hu, J., Li, X., Wu, C., Yang, C., Hua, H., and Gao, G. (2012). Pyramiding and evaluation of the brown planthopper resistance genes *Bph14* and *Bph15* in hybrid rice. *Mol. Breed.* 29, 61–69. doi: 10.1007/s11032-010-9526-x

Hu, J., Xiao, C., and He, Y. (2016). Recent progress on the genetics and molecular breeding of brown planthopper resistance in rice. *Rice* 9 (1), 1–12. doi: 10.1186/s12284-016-0099-0

Hua, L., Wu, J., Chen, C., Wu, W., He, X., Lin, F., et al. (2012). The isolation of *Pi1*, an allele at the *Pik* locus which confers broad spectrum resistance to rice blast. *Theoret. Appl. Genet.* 125, 1047–1055. doi: 10.1007/s00122-012-1894-7

Hutin, M., Sabot, F., Ghesquière, A., Koebnik, R., and Szurek, B. (2015). A knowledge-based molecular screen uncovers a broad-spectrum *Os SWEET 14* resistance allele to bacterial blight from wild rice. *Plant J.* 84 (4), 694–703. doi: 10.1111/tpj.13042

Inoue, H., Nakamura, M., Mizubayashi, T., Takahashi, A., Sugano, S., Fukuoka, S., et al. (2017). Panicle blast 1 (*Pb1*) resistance is dependent on at least four QTLs in the rice genome. *Rice* 10 (1), 1–10. doi: 10.1186/s12284-017-0175-0

Inukai, T., Nagashima, S., and Kato, M. (2019). *Pid3-I1* is a race-specific partial-resistance allele at the *Pid3* blast resistance locus in rice. *Theor. Appl. Genet.* 132, 395–404. doi: 10.1007/s00122-018-3227-y

IRRI Rice Knowledge Bank. Available at: http://www.knowledgebank.irri.org/step-by-step-production/growth/pests-and-diseases.

Iyer, A. S., and McCouch, S. R. (2004). The rice bacterial blight resistance gene *xa5* encodes a novel form of disease resistance. *Mol. Plant Microbe Interact.* 17, 1348–1354. doi: 10.1094/MPMI.2004.17.12.1348

Jairin, J., Phengrat, K., Teangdeerith, S., Vanavichit, A., and Toojinda, T. (2007). Mapping of a broad-spectrum brown planthopper resistance gene, *Bph3*, on rice chromosome 6. *Mol. Breed.* 19, 35–44. doi: 10.1007/s11032-006-9040-3

Jamaloddin, M., Mahender, A., Gokulan, C. G., BalachIranjeevi, C., Maliha, A., Patel, H. K., et al. (2021). "Molecular approaches for disease resistance in rice", *Rice Improvement*, eds. J. Ali and S. H. Wani (Cham: Springer) 315-378. doi: 10.1007/978-3-030-66530-2_10

Jena, K. K., Hechanova, S. L., Verdeprado, H., Prahalada, G. D., and Kim, S. R. (2017). Development of 25 near-isogenic lines (NILs) with ten BPH resistance genes in rice (*Oryza sativa* L.): production, resistance spectrum, and molecular analysis. *Theor. Appl. Genet.* 130, 2345–2360. doi: 10.1007/s00122-017-2963-8

Jena, K. K., and Kim, S. M. (2010). Current status of brown planthopper (BPH) resistance and genetics. *Rice* 3, 161–171. doi: 10.1007/s12284-010-9050-y

Jeung, J. U., Heu, S. G., Shin, M. S., Cruz, C. M., and Jena, K. K. (2006). Dynamics of *Xanthomonas oryzae* pv. *oryzae* populations in Korea and their relationship to known bacterial blight resistance genes. *Phytopathology* 96, 867–875. doi: 10.1094/PHYTO-96-0867

Jeung, J. U., Kim, B. R., Cho, Y. C., Han, S. S., Moon, H. P., Lee, Y. T., et al. (2007). A novel gene, Pi40 (t), linked to the DNA markers derived from NBS-LRR motifs confers broad spectrum of blast resistance in rice. *Theor. Appl. Genet.* 115, 1163–1177. doi: 10.1007/s00122-007-0642-x

Ji, C., Ji, Z., Liu, B., Cheng, H., Liu, H., Liu, S., et al. (2020). *Xa1* allelic R genes activate rice blight resistance suppressed by interfering TAL effectors. *Plant Commun.* 1 (4), 100087. doi: 10.1016/j.xplc.2020.100087

Ji, H., Kim, S. R., Kim, Y. H., Suh, J. P., Park, H. M., Sreenivasulu, N., et al. (2016). Map-based cloning and characterization of the *BPH18* gene from wild rice conferring resistance to brown planthopper (BPH) insect pest. *Sci. Rep.* 6 (1), 1–14. doi: 10.1038/srep34376

Jiang, H., Li, Z., Liu, J., Shen, Z., Gao, G., Zhang, Q., et al. (2019). Development and evaluation of improved lines with broad-spectrum resistance to rice blast using nine resistance genes. *Rice* 12 (1), 29. doi: 10.1186/s12284-019-0292-z

Jiang, N., Yan, J., Liang, Y., Shi, Y., He, Z., Wu, Y., et al. (2020). Resistance genes and their interactions with bacterial blight/leaf streak pathogens (*Xanthomonas oryzae*) in rice (*Oryza sativa* L.)—an updated review. *Rice* 13 (1), 1–12. doi: 10.1186/s12284-019-0358-y

Kabish, A., and Khush, G. S. (1988). Genetic analysis of resistance to brown planthopper in rice (*Oryza sativa* L.). *Plant Breed.* 100, 54–58. doi: 10.1111/j.1439-0523.1988.tb00216.x

Ke, Y., Deng, H., and Wang, S. (2017). Advances in understanding broad-spectrum resistance to pathogens in rice. *Plant J.* 90 (4), 738–748. doi: 10.1111/tpj.13438

Khush, G. S. (2005). What it will take to feed 5.0 billion rice consumers in 2030. *Plant Mol. Biol.* 59, 1–6. doi: 10.1007/s11103-005-2159-5

Kim, S. R., Ramos, J., Ashikari, M., Virk, P. S., Torres, E. A., Nissila, E., et al. (2016). Development and validation of allele-specific SNP/indel markers for eight yield-enhancing genes using whole-genome sequencing strategy to increase yield potential of rice, *Oryza sativa* L. *Rice* 9 (1), 1–17. doi: 10.1186/s12284-016-0084-7

Kim, E. G., Yun, S., Park, J. R., and Kim, K. M. (2021). Identification of F3H, major secondary metabolite-related gene that confers resistance against whitebacked planthopper through QTL mapping in rice. *Plants* 10 (1), 81. doi: 10.3390/plants10010081

Kouassi, N. K., N'guessan, P., Albar, L., Fauquet, C. M., and Brugidou, C. (2005). Distribution and characterization of Rice yellow mottle virus: a threat to African farmers. *Plant Dis.* 89 (2), 124–133. doi: 10.1094/PD-89-0124

Kuang, Y.-H., Fang, Y.-F., Lin, S.-C., Tsai, S.-F., Yang, Z.-W., Li, C.-P., et al. (2021). The impact of climate change on the resistance of rice near-isogenic lines with resistance genes against brown planthopper. *Rice* 14, 64. doi: 10.1186/s12284-021-00508-6

Lanoiselet, V. M., Cother, E. J., and Ash, G. J. (2007). Aggregate sheath spot and sheath spot of rice. *Crop Prot.* 26, 799–808. doi: 10.1016/j.cropro.2006.06.016

Lee, S. B., Kim, N., Jo, S., Hur, Y. J., Lee, J. Y., Cho, J. H., et al. (2021). Mapping of a major QTL, qBK1Z, for bakanae disease resistance in rice. *Plants* 10 (3), 434. doi: 10.3390/plants10030434

Lee, J. H., Muhsin, M., Atienza, G. A., Kwak, D. Y., Kim, S. M., De Leon, T. B., et al. (2010). Single nucleotide polymorphisms in a gene for translation initiation factor (*eIF4G*) of rice (*Oryza sativa*) associated with resistance to Rice tungro spherical virus. *Mol. Plant Microbe Interact.* 23 (1), 29–38. doi: 10.1094/MPMI-23-1-0029

Lee, S. K., Song, M. Y., Seo, Y. S., Kim, H. K., Ko, S., Cao, P. J., et al. (2009). Rice *Pi5*-mediated resistance to *Magnaporthe oryzae* requires the presence of two coiled-coil–nucleotide-binding–leucine-rich repeat genes. *Genetics* 181 (4), 1627–1638. doi: 10.1534/genetics.108.099226

Leelagud, P., Kongsila, S., Vejchasarn, P., Darwell, K., Phansenee, Y., Suthanthangjai, A., et al. (2020). Genetic diversity of Asian rice gall midge based on *mtCOI* gene sequences and identification of a novel resistance locus *gm12* in rice cultivar MN62M. *Mol. Biol. Rep.* 47, 4273–4283. doi: 10.1007/s11033-020-05546-9

Li, W., Chern, M., Yin, J., Wang, J., and Chen, X. (2019). Recent advances in broad-spectrum resistance to the rice blast disease. *Curr. Opin. Plant Biol.* 50, 114–120. doi: 10.1016/j.pbi.2019.03.015

Li, W., Zhu, Z., Chern, M., Yin, J., Yang, C., Ran, L., et al. (2017). A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell* 170 (1), 114–126. doi: 10.1016/j.cell.2017.06.008

Lin, F., Chen, S., Que, Z., Wang, L., Liu, X., and Pan, Q. (2007). The blast resistance gene *Pi37* encodes a nucleotide binding site–leucine-rich repeat protein and is a member of a resistance gene cluster on rice chromosome 1. *Genetics* 177 (3), 1871–1880. doi: 10.1534/genetics.107.080648

Lin, S.-C., Li, Y., Hu, F.-Y., Wang, C.-L., Kuang, Y.-H., Sung, C.-L., et al. (2022). Effect of nitrogen fertilizer on the resistance of rice near-isogenic lines with BPH resistance genes. *Bot. Stud.* 63, 16. doi: 10.1186/s40529-022-00347-8

Liu, X., Lin, F., Wang, L., and Pan, Q. (2007). The in *silico* map-based cloning of *Pi36*, a rice coiled-coil–nucleotide-binding site–leucine-rich repeat gene that confers race-specific resistance to the blast fungus. *Genetics* 176 (4), 2541–2549. doi: 10.1534/genetics.107.075465

Liu, W., Liu, J., Ning, Y., Ding, B., Wang, X., Wang, Z., et al. (2013). Recent progress in understanding PAMP-and effector-triggered immunity against the rice blast fungus *Magnaporthe oryzae*. *Mol. Plant* 6 (3), 605–620. doi: 10.1093/mp/sst015

Liu, Y., Liu, B., Zhu, X., Yang, J., Bordeos, A., Wang, G., et al. (2013). Fine-mapping and molecular marker development for Pi56 (t), a NBS-LRR gene conferring broad-spectrum resistance to *Magnaporthe oryzae* in rice. *Theoret. Appl. Genet.* 126, 985–998. doi: 10.1007/s00122-012-2031-3

Liu, Y., Wu, H., Chen, H., Liu, Y., He, J., Kang, H., et al. (2015). A gene cluster encoding lectin receptor kinases confers broad-spectrum and durable insect resistance in rice. *Nat. Biotechnol.* 33 (3), 301–305. doi: 10.1038/nbt.3069

Liu, Q., Yuan, M., Zhou, Y. A. N., Li, X., Xiao, J., and Wang, S. (2011). A paralog of the MtN3/saliva family recessively confers race-specific resistance to *Xanthomonas oryzae* in rice. *Plant Cell Environ.* 34 (11), 1958–1969. doi: 10.1111/j.1365-3040.2011.02391.x

Liu, Z., Zhu, Y., Shi, H., Qiu, J., Ding, X., and Kou, Y. (2021). Recent progress in rice broad-spectrum disease resistance. *Int. J. Mol. Sci.* 22 (21), 11658. doi: 10.3390/ijms222111658

Ma, J., Lei, C., Xu, X., Hao, K., Wang, J., Cheng, Z., et al. (2015). *Pi64*, encoding a novel CC-NBS-LRR protein confers resistance to leaf and neck blast in rice. *Mol. Plant Microbe Interact.* 28 (5), 558–568. doi: 10.1094/MPMI-11-14-0367-R

Ma, L., Yu, Y., Li, C., Wang, P., Liu, K., Ma, W., et al. (2022). Genome-wide association study identifies a rice panicle blast resistance gene *Pb3* encoding NLR protein. *Int. J. Mol. Sci.* 23, 14032. doi: 10.3390/ijms232214032

Macovei, A., Sevilla, N. R., Cantos, C., Jonson, G. B., Slamet-Loedin, I., Čermák, T., et al. (2018). Novel alleles of rice *eIF4G* generated by CRISPR/Cas9-targeted mutagenesis confer resistance to Rice tungro spherical virus. *Plant Biotechnol. J.* 16 (11), 1918–1927. doi: 10.1111/pbi.12927

Matsumoto, K., Ota, Y., Seta, S., Nakayama, Y., Ohno, T., Mizobuchi, R., et al. (2017). Identification of QTLs for rice brown spot resistance in backcross inbred lines derived from a cross between Koshihikari and CH45. *Breed. Sci.* 67 (5), 540–543. doi: 10.1270/jsbbs.17057

McCouch, S. R., Teytelman, L., Xu, Y., Lobos, K. B., Clare, K., Walton, M., et al. (2002). Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res.* 9 (6), 199–207. doi: 10.1093/dnares/9.6.199

McKenzie, K. S., Johnson, C. W., Tseng, S. T., Oster, J. J., and Brandon, D. M. (1994). Breeding improved rice cultivars for temperate regions: a case study. *Aust. J. Exp. Agric.* 34 (7), 897–905. doi: 10.1071/EA9940897

Meng, X., Xiao, G., Telebanco-Yanoria, M. J., Siazon, P. M., Padilla, J., Opulencia, R., et al. (2020). The broad-spectrum rice blast resistance (R) gene *Pita2* encodes a novel R protein unique from *Pita*. *Rice* 13 (19). doi: 10.1186/s12284-020-00377-5

Mizobuchi, R., Fukuoka, S., Tsushima, S., Yano, M., and Sato, H. (2016). QTLs for resistance to major rice diseases exacerbated by global warming: brown spot, bacterial seedling rot, and bacterial grain rot. *Rice* 9 (1), 1–12. doi: 10.1186/s12284-016-0095-4

Molla, K. A., Karmakar, S., Molla, J., Bajaj, P., Varshney, R. K., Datta,, et al. (2020). Understanding sheath blight resistance in rice: the road behind and the road ahead. *Plant Biotechnol. J.* 18 (4), 895–915. doi: 10.1111/pbi.13312

Musonerimana, S., Bez, C., Licastro, D., Habarugira, G., Bigirimana, J., and Venturi, V. (2020). Pathobiomes revealed that *Pseudomonas fuscovaginae* and *Sarocladium oryzae* are independently associated with rice. *Microb. Ecol.* 80, 627–642. doi: 10.1007/s00248-020-01529-2

Nadeem, M. A., Nawaz, M. A., Shahid, M. Q., Doğan, Y., Comertpay, G., and Yıldız, M. (2018). DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. *Biotechnol. Biotechnol. Equip.* 32 (2), 261–285. doi: 10.1080/13102818.2017.1400401

Naik, S. B., Divya, D., Sahu, N., Sundaram, R. M., Sarao, P. S., Singh, K., et al. (2018). A new gene *Bph33* (t) conferring resistance to brown planthopper (BPH), *Nilaparvata lugens* (Stål) in rice line RP2068-18-3-5. *Euphytica* 214, 1–12. doi: 10.1007/s10681-018-2131-5

Ndjiondjop, M. N., Albar, L., Fargette, D., Fauquet, C., and Ghesquière, A. (1999). The genetic basis of high resistance to rice yellow mottle virus (RYMV) in cultivars of two cultivated rice species. *Plant Dis.* 83 (10), 931–935. doi: 10.1094/PDIS.1999.83.10.931

Neelam, K., Kumar, K., Kaur, A., Kishore, A., Kaur, P., Babbar, A., et al. (2022). High-resolution mapping of the quantitative trait locus (QTLs) conferring resistance to false smut disease in rice. *J. Appl. Genet.* 63 (1), 35–45. doi: 10.1007/s13353-021-00659-8

Ogawa, T., Yamamoto, T., Khush, G. S., and Mew, T. W. (1991). Breeding of near-isogenic lines of rice with single genes for resistance to bacterial blight pathogen (*Xanthomonas campestris* pv. *oryzae*). *Jpn. J. Breed.* 41 (3), 523–529. doi: 10.1270/jsbbs1951.41.523

Okuyama, Y., Kanzaki, H., Abe, A., Yoshida, K., Tamiru, M., Saitoh, H., et al. (2011). A multifaceted genomics approach allows the isolation of the rice *Pia*-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *Plant J.* 66 (3), 467–479. doi: 10.1111/j.1365-313X.2011.04502.x

Oliva, R., Ji, C., Atienza-Grande, G., Huguet-Tapia, J. C., Perez-Quintero, A., Li, T., et al. (2019). Broad-spectrum resistance to bacterial blight in rice using genome editing. *Nat. Biotechnol.* 37 (11), 1344–1350. doi: 10.1038/s41587-019-0267-z

Orjuela, J., Deless, E. T., Kolade, O., Chéron, S., Ghesquière, A., and Albar, L. (2013). A recessive resistance to rice yellow mottle virus is associated with a rice homolog of the *CPR5* gene, a regulator of active defense mechanisms. *Mol. Plant Microbe Interact.* 26 (12), 1455–1463. doi: 10.1094/MPMI-05-13-0127-R

Ortega, L., and Rojas, C. M. (2021). Bacterial Panicle Blight and *Burkholderia glumae*: From pathogen biology to disease control. *Phytopathology* 111 (5), 772–778. doi: 10.1094/PHYTO-09-20-0401-RVW

Oster, J. J. (1992). Reaction of a resistant breeding line and susceptible California rice cultivars to *Sclerotium oryzae*. *Plant Dis.* 76 (7), 740–744. doi: 10.1094/PD-76-0740

Pathak, M. D., and Khan, Z. R. (1994). *Insect Pests of Rice* (Philippines: International Rice Research Institute).

Pennisi, E. (2010). Armed and dangerous. *Science* 327, 804–805. doi: 10.1126/science.327.5967.804

Phi, C. N., Fujita, D., Yamagata, Y., Yoshimura, A., and Yasui, H. (2019). High-resolution mapping of *GRH6*, a gene from *Oryza nivara* (Sharma et Shastry) conferring resistance to green rice leafhopper (*Nephotettix cincticeps* Uhler). *Breed. Sci.* 69 (3), 439–446. doi: 10.1270/jsbbs.19029

Pidon, H., Chéron, S., Ghesquière, A., and Albar, L. (2020). Allele mining unlocks the identification of RYMV resistance genes and alleles in African cultivated rice. *BMC Plant Biol.* 20 (1), 1–14. doi: 10.1186/s12870-020-02433-0

Pidon, H., Ghesquière, A., Chéron, S., Issaka, S., Hébrard, E., Sabot, F., et al. (2017). Fine mapping of *RYMV3*: a new resistance gene to Rice yellow mottle virus from Oryza glaberrima. *Theor. Appl. Genet.* 130, 807–818. doi: 10.1007/s00122-017-2853-0

Pradhan, S. K., Nayak, D. K., Mohanty, S., Behera, L., Barik, S. R., Pandit, E., et al. (2015). Pyramiding of three bacterial blight resistance genes for broad-spectrum resistance in deepwater rice variety, Jalmagna. *Rice* 8, 19. doi: 10.1186/s12284-015-0051-8

Pradhan, S. K., Barik, S. R., Nayak, D. K., Pradhan, A., Pandit, E., Nayak,, et al. (2020). Genetics, molecular mechanisms and deployment of bacterial blight resistance genes in rice. *Crit. Rev. Plant Sci.* 39 (4), 360–385. doi: 10.1080/07352689.2020.1801559

Qin, J., Wang, C., Wang, L., Zhao, S., and Wu, J. (2019). Defense and counter-defense in rice–virus interactions. *Phytopathol. Res.* 1, 1–6. doi: 10.1186/s42483-019-0041-7

Qiu, Y., Guo, J., Jing, S., Zhu, L., and He, G. (2012). Development and characterization of *japonica* rice lines carrying the brown planthopper-resistance genes *BPH12* and *BPH6*. *Theor. Appl. Genet.* 124, 485–494. doi: 10.1007/s00122-011-1722-5

Qu, S., Liu, G., Zhou, B., Bellizzi, M., Zeng, L., Dai, L., et al. (2006). The broad-spectrum blast resistance gene *Pi9* encodes a nucleotide-binding site–leucine-rich repeat protein and is a member of a multigene family in rice. *Genetics* 172 (3), 1901–1914. doi: 10.1534/genetics.105.044891

Rao, Y., Dong, G., Zeng, D., Hu, J., Zeng, L., Gao, Z., et al. (2010). Genetic analysis of leaffolder resistance in rice. *J. Genet. Genomics* 37 (5), 325–331. doi: 10.1016/S1673-8527(09)60050-3

Ren, J., Gao, F., Wu, X., Lu, X., Zeng, L., Lv, J., et al. (2016). *Bph32*, a novel gene encoding an unknown SCR domain-containing protein, confers resistance against the brown planthopper in rice. *Sci. Rep.* 6 (1), 37645. doi: 10.1038/srep37645

Rosas, J. E., Martínez, S., Blanco, P., Perez de Vida, F., Bonnecarrère, V., Mosquera, G., et al. (2018). Resistance to multiple temperate and tropical stem and sheath diseases of rice. *Plant Genome* 11 (1), 170029. doi: 10.3835/plantgenome2017.03.0029

Sama, V. S. A. K., Rawat, N., Sundaram, R. M., Himabindu, K., Naik, B. S., Viraktamath, B. C., et al. (2014). A putative candidate for the recessive gall midge resistance gene *gm3* in rice identified and validated. *Theor. Appl. Genet.* 127, 113–124. doi: 10.1007/s00122-013-2205-7

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 3 (3), 430–439. doi: 10.1038/s41559-018-0793-y

Schneider, P., and Asch, F. (2020). Rice production and food security in Asian Mega deltas—A review on characteristics, vulnerabilities and agricultural adaptation options to cope with climate change. *J. Agron. Crop Sci.* 206 (4), 491–503. doi: 10.1111/jac.12415

Seck, P. A., Diagne, A., Mohanty, S., and Wopereis, M. C. (2012). Crops that feed the world 7: Rice. *Food Sec.* 4, 7–24. doi: 10.1007/s12571-012-0168-1

Shang, J., Tao, Y., Chen, X., Zou, Y., Lei, C., Wang, J., et al. (2009). Identification of a new rice blast resistance gene, *Pid3*, by genome-wide comparison of paired nucleotide-binding site–leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes. *Genetics* 182 (4), 1303–1311. doi: 10.1534/genetics.109.102871

Shi, S., Wang, H., Nie, L., Tan, D. I., Zhou, C., Zhang, Q., et al. (2021). *Bph30* confers resistance to brown planthopper by fortifying sclerenchyma in rice leaf sheaths. *Mol. Plant* 14, 1714–1732. doi: 10.1016/j.molp.2021.07.004

Si, W., Yuan, Y., Huang, J., Zhang, X., Zhang, Y., Zhang, Y., et al. (2015). Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol.* 206 (4), 1491–1502. doi: 10.1111/nph.13319

Singh, P., Verma, R. L., Singh, R. S., Singh, R. P., Singh, H. B., Arsode, P., et al. (2020). "Biotic stress management in rice (*Oryza sativa* L.) through conventional and molecular approaches," in *New Frontiers in Stress Management For Durable Agriculture*. Eds. A. Rakshit, H. Singh, A. Singh, U. Singh and L. Fraceto (Singapore: Springer), 609–644.

Song, W. Y., Wang, G. L., Chen, L. L., Kim, H. S., Pi, L. Y., Holsten, T., et al. (1995). A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* 270 (5243), 1804–1806. doi: 10.1126/science.270.5243.1804

Su, J., Wang, W., Han, J., Chen, S., Wang, C., Zeng, L., et al. (2015). Functional divergence of duplicated genes results in a novel blast resistance gene *Pi50* at the *Pi2/9* locus. *Theoret. Appl. Genet.* 128, 2213–2225. doi: 10.1007/s00122-015-2579-9

Sun, X., Cao, Y., Yang, Z., Xu, C., Li, X., Wang, S., et al. (2004). *Xa26*, a gene conferring resistance to *Xanthomonas oryzae* pv. *oryzae* in rice, encodes an LRR receptor kinase-like protein. *Plant J.* 37 (4), 517–527. doi: 10.1046/j.1365-313X.2003.01976.x

Sun, L., Su, C., Wang, C., Zhai, H., and Wan, J. (2005). Mapping of a major resistance gene to the brown planthopper in the rice cultivar Rathu Heenati. *Breed. Sci.* 55 (4), 391–396. doi: 10.1270/jsbbs.55.391

Takahashi, A., Hayashi, N., Miyao, A., and Hirochika, H. (2010). Unique features of the rice blast resistance *Pish* locus revealed by large scale retrotransposon-tagging. *BMC Plant Biol.* 10, 1–14. doi: 10.1186/1471-2229-10-175

Tamura, Y., Hattori, M., Yoshioka, H., Yoshioka, M., Takahashi, A., and Wu, J. (2014). Map-based cloning and characterization of a brown planthopper resistance gene *BPH26* from *Oryza sativa* L. ssp. *indica* cultivar ADR52. *Sci. Rep.* 4 (1), 1–8. doi: 10.1038/srep05872

Tang, D., Wu, W., Li, W., Lu, H., and Worland, A. J. (2000). Mapping of QTLs conferring resistance to bacterial leaf streak in rice. *Theor. Appl. Genet.* 101, 286–291. doi: 10.1007/s001220051481

Telebanco-Yanoria, M. J., Koide, Y., Fukuta, Y., Imbe, T., Kato, H., Tsunematsu, H., et al. (2010). Development of near-isogenic lines of *Japonica*-type rice variety Lijiangxintuanheigu as differentials for blast resistance. *Breed. Sci.* 60 (5), 629–638. doi: 10.1270/jsbbs.60.629

Telebanco-Yanoria, M. J., Koide, Y., Fukuta, Y., Imbe, T., Tsunematsu, H., Kato, H., et al. (2011). A set of near-isogenic lines of *Indica*-type rice variety CO 39 as differential varieties for blast resistance. *Mol. Breed.* 27, 357–373. doi: 10.1007/s11032-010-9437-x

Thiémélé, D., Boisnard, A., Ndjiondjop, M. N., Chéron, S., Séré, Y., Aké, S., et al. (2010). Identification of a second major resistance gene to Rice yellow mottle virus, *RYMV2*, in the African cultivated rice species, *O. glaberrima*. *Theor. Appl. Genet.* 121, 169–179. doi: 10.1007/s00122-010-1300-2

Tian, D., Wang, J., Zeng, X., Gu, K., Qiu, C., and Yang, X. (2014). The rice TAL effector–dependent resistance protein *XA10* triggers cell death and calcium depletion in the endoplasmic reticulum. *Plant Cell* 26 (1), 497–515. doi: 10.1105/tpc.113.119255

Velásquez, A. C., Castroverde, C. D. M., and He, S. Y. (2018). Plant–pathogen warfare under changing climate conditions. *Curr. Biol.* 28 (10), R619–R634. doi: 10.1016/j.cub.2018.03.054

Wang, Y., Cao, L., Zhang, Y., Cao, C., Liu, F., Huang, F., et al. (2015). Map-based cloning and characterization of *BPH29*, a B3 domain-containing recessive gene conferring brown planthopper resistance in rice. *J. Exp. Bot.* 66 (19), 6035–6045. doi: 10.1093/jxb/erv318

Wang, C., Chen, S., Feng, A., Su, J., Wang, W., Feng, J., et al. (2021). *Xa7*, a small orphan gene harboring promoter trap for *AvrXa7*, leads to the durable resistance to *Xanthomonas oryzae* pv. *oryzae*. *Rice* 14 (1), 1–16. doi: 10.1186/s12284-021-00490-z

Wang, H. M., Chen, J., Shi, Y. F., Pan, G., Shen, H. C., and Wu, J. L. (2012). Development and validation of CAPS markers for marker-assisted selection of rice blast resistance gene *Pi25*. *Acta Agron. Sin.* 38 (11), 1960–1968. doi: 10.3724/SP.J.1006.2012.01960

Wang, Q., Liu, Y., He, J., Zheng, X., Hu, J., Liu, Y., et al. (2014). *STV11* encodes a sulphotransferase and confers durable resistance to rice stripe virus. *Nat. Commun.* 5, 4768. doi: 10.1038/ncomms5768

Wang, Z. X., Yano, M., Yamanouchi, U., Iwamoto, M., Monna, L., Hayasaka, H., et al. (1999). The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J.* 19 (1), 55–64. doi: 10.1046/j.1365-313X.1999.00498.x

Wang, C., Zhang, X., Fan, Y., Gao, Y., Zhu, Q., Zheng, C., et al. (2015). *XA23* is an executor R protein and confers broad-spectrum disease resistance in rice. *Mol. Plant* 8 (2), 290–302. doi: 10.1016/j.molp.2014.10.010

Webb, K. M., Ona, I., Bai, J., Garrett, K. A., Mew, T., Cruz, C. M. V., et al. (2010). A benefit of high temperature: increased effectiveness of a rice bacterial blight disease resistance gene. *New Phytol.* 185, 568–576. doi: 10.1111/j.1469-8137.2009.03076.x

Wu, Y., Xiao, N., Chen, Y., Yu, L., Pan, C., Li, Y., et al. (2019). Comprehensive evaluation of resistance effects of pyramiding lines with different broad-spectrum resistance genes against *Magnaporthe oryzae* in rice (*Oryza sativa* L.). *Rice* 12, 1–13. doi: 10.1186/s12284-019-0264-3

Xie, X., Chen, Z., Cao, J., Guan, H., Lin, D., Li, C., et al. (2014). Toward the positional cloning of *qBlsr5a*, a QTL underlying resistance to bacterial leaf streak, using overlapping sub-CSSLs in rice. *PloS One* 9 (4), e95751. doi: 10.1371/Journal.Pone.0095751

Xie, Z., Yan, B., Shou, J., Tang, J., Wang, X., Zhai, K., et al. (2019). A nucleotide-binding site-leucine-rich repeat receptor pair confers broad-spectrum disease resistance through physical association in rice. *Philos. Trans. R. Soc B.* 374 (1767), 20180308. doi: 10.1098/rstb.2018.0308

Xie, X., Zheng, Y., Lu, L., Yuan, J., Hu, J., Bu, S., et al. (2021). Genome-wide association study of QTLs conferring resistance to bacterial leaf streak in rice. *Plants* 10 (10), 2039. doi: 10.3390/plants10102039

Xing, J., Zhang, D., Yin, F., Zhong, Q., Wang, B., Xiao, S., et al. (2021). Identification and fine-mapping of a new bacterial blight resistance gene, Xa47 (t), in G252, an introgression line of Yuanjiang common wild rice (*Oryza rufipogon*). *Plant Dis.* 105 (12), 4106–4112. doi: 10.1094/PDIS-05-21-0939-RE

Xu, X., Hayashi, N., Wang, C. T., Fukuoka, S., Kawasaki, S., Takatsuji, H., et al. (2014). Rice blast resistance gene Pikahei-1 (t), a member of a resistance gene cluster on chromosome 4, encodes a nucleotide-binding site and leucine-rich repeat protein. *Mol. Breed.* 34, 691–700. doi: 10.1007/s11032-014-0067-6

Yan, L., Luo, T., Huang, D., Wei, M., Ma, Z., Liu, C., et al. (2023). Recent advances in molecular mechanism and breeding utilization of brown planthopper resistance genes in rice: An integrated review. *Int. J. Mol. Sci.* 24 (15), 12061. doi: 10.3390/ijms241512061

Yang, M., Lin, J., Cheng, L., Zhou, H., Chen, S., Liu, F., et al. (2020). Identification of a novel planthopper resistance gene from wild rice (*Oryza rufipogon* Griff.). *Crop J.* 8 (6), 1057–1070. doi: 10.1016/j.cj.2020.03.011

Yoshimura, S., Yamanouchi, U., Katayose, Y., Toki, S., Wang, Z. X., Kono, I., et al. (1998). Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl. Acad. Sci.* 95 (4), 1663–1668. doi: 10.1073/pnas.95.4.1663

Yu, Y., Ma, L., Wang, X., Zhao, Z., Wang, W., Fan, Y., et al. (2022). Genome-wide association study identifies a rice panicle blast resistance gene, *Pb2*, encoding. *Int. J. Mol. Sci.* 23 (10), 5668. doi: 10.3390/ijms23105668

Yuan, B., Zhai, C., Wang, W., Zeng, X., Xu, X., Hu, H., et al. (2011). The *Pik-p* resistance to *Magnaporthe oryzae* in rice is mediated by a pair of closely linked CC-NBS-LRR genes. *Theoret. Appl. Genet.* 122, 1017–1028. doi: 10.1007/s00122-010-1506-3

Zampieri, E., Volante, A., Marè, C., Orasen, G., Desiderio, F., Biselli, C., et al. (2023). Marker-assisted pyramiding of blast-resistance genes in a *japonica* elite rice cultivar through forward and background selection. *Plants* 12, 757. doi: 10.3390/plants12040757

Zarbafi, S. S., and Ham, J. H. (2019). An overview of rice QTLs associated with disease resistance to three major rice diseases: blast, sheath blight, and bacterial panicle blight. *Agronomy* 9 (4), 177. doi: 10.3390/agronomy9040177

Zhai, C., Lin, F., Dong, Z., He, X., Yuan, B., Zeng, X., et al. (2011). The isolation and characterization of *Pik*, a rice blast resistance gene which emerged after rice domestication. *New Phytol.* 189 (1), 321–334. doi: 10.1111/j.1469-8137.2010.03462.x

Zhang, L., Nakagomi, Y., Endo, T., Teranishi, M., Hidema, J., Sato, S., et al. (2018). Divergent evolution of rice blast resistance *Pi54* locus in the genus *Oryza*. *Rice* 11, 1–13. doi: 10.1186/s12284-018-0256-8

Zhao, Y., Huang, J., Wang, Z., Jing, S., Wang, Y., Ouyang, Y., et al. (2016). Allelic diversity in an NLR gene *BPH9* enables rice to combat planthopper variation. *Proc. Natl. Acad. Sci.* 113 (45), 12850–12855. doi: 10.1073/pnas.1614862113

Zhao, H., Wang, X., Jia, Y., Minkenberg, B., Wheatley, M., Fan, J., et al. (2018). The rice blast resistance gene *Ptr* encodes an atypical protein required for broad-spectrum disease resistance. *Nat. Commun.* 9 (1) 2039. doi: 10.1038/s41467-018-04369-4

Zhou, X. G. (2019). "Sustainable strategies for managing bacterial panicle blight in rice," in *Protecting Rice Grains in The Post-Genomic Era*. Ed. Y. Jia (London, UK: IntechOpen), 67–80.

Zhou, B., Qu, S., Liu, G., Dolan, M., Sakai, H., Lu, G., et al. (2006). The eight amino-acid differences within three leucine-rich repeats between *Pi2* and *Piz-t* resistance proteins determine the resistance specificity to *Magnaporthe grisea*. *Mol. Plant Microbe Interact.* 19 (11), 1216–1228. doi: 10.1094/MPMI-19-1216

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 October 2023
DOI 10.3389/fpls.2023.1265574

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Hongjian Zheng,
Shanghai Academy of Agricultural Sciences, China
Toi J. Tsilo,
Agricultural Research Council of South Africa (ARC-SA), South Africa

*CORRESPONDENCE
Wenbin Wang
lnsoybean@163.com
Shuhong Song
songshuhong2017@163.com

RECEIVED 23 July 2023
ACCEPTED 18 September 2023
PUBLISHED 09 October 2023

CITATION
Li S, Cao Y, Wang C, Yan C, Sun X, Zhang L, Wang W and Song S (2023) Genome-wide association mapping for yield-related traits in soybean (Glycine max) under well-watered and drought-stressed conditions. *Front. Plant Sci.* 14:1265574. doi: 10.3389/fpls.2023.1265574

Genome-wide association mapping for yield-related traits in soybean (Glycine max) under well-watered and drought-stressed conditions

Shengyou Li, Yongqiang Cao, Changling Wang, Chunjuan Yan, Xugang Sun, Lijun Zhang, Wenbin Wang* and Shuhong Song*

Institute of Crop Research, Liaoning Academy of Agricultural Sciences, Shenyang, China

Soybean (*Glycine max*) productivity is significantly reduced by drought stress. Breeders are aiming to improve soybean grain yields both under well-watered (WW) and drought-stressed (DS) conditions, however, little is known about the genetic architecture of yield-related traits. Here, a panel of 188 soybean germplasm was used in a genome wide association study (GWAS) to identify single nucleotide polymorphism (SNP) markers linked to yield-related traits including pod number per plant (PN), biomass per plant (BM) and seed weight per plant (SW). The SLAF-seq genotyping was conducted on the population and three phenotype traits were examined in WW and DS conditions in four environments. Based on best linear unbiased prediction (BLUP) data and individual environmental analyses, 39 SNPs were significantly associated with three soybean traits under two conditions, which were tagged to 26 genomic regions by linkage disequilibrium (LD) analysis. Of these, six QTLs qPN-WW19.1, qPN-DS8.8, qBM-WW1, qBM-DS17.4, qSW-WW4 and qSW-DS8 were identified controlling PN, BM and SW of soybean. There were larger proportions of favorable haplotypes for locus qPN-WW19.1 and qSW-WW4 rather than qBM-WW1, qBM-DS17.4, qPN-DS8.8 and qSW-DS8 in both landraces and improved cultivars. In addition, several putative candidate genes such as *Glyma.19G211300*, *Glyma.17G057100* and *Glyma.04G124800*, encoding E3 ubiquitin-protein ligase BAH1, WRKY transcription factor 11 and protein zinc induced facilitator-like 1, respectively, were predicted. We propose that the further exploration of these locus will facilitate accelerating breeding for high-yield soybean cultivars.

KEYWORDS

drought stress, favorable haplotypes, GWAS, soybean (Glycine max), yield-related traits

1 Introduction

Soybean (*Glycine max*) is known as the main source of plant oil and protein in the world (Cerezini et al., 2016). However, the sustainability of soybean production is threatened by persistent droughts with the climatic changes (Chen et al., 2016). Field and greenhouse experiments have shown significant reduction of 24-50% in soybean grain yield by drought stress (Frederick et al., 2001). Reduction of grain yield is maximal while water deficiency happens during flowering and podding stage, which is due to decreases in pod number per plant (PN), biomass per plant (BM) and seed weight per plant (SW) in soybean. Due to carbohydrate deprivation, drought-induced lower photosynthetic capacity increased pod abortion and decreased dry matter production after anthesis (Liu et al., 2004). Thus, Breeding for new soybean cultivars with high SW as well as PN and BM both under well-watered and drought-stressed conditions is therefore an important strategy for addressing this imminent threat to food security.

Selecting genotypes with better genetic gains in soybean can improve the efficiency of cultivar development programs based on genomic information of these yield-related traits (Yoosefzadeh Najafabadi, 2021). The traditional QTL linkage mapping of pod number per plant (PN) (Sun et al., 2022), biomass per plant (BM) (Yang, 2021), and seed weight per plant (SW) (Hacisalihoglu et al., 2018) in soybean, has made some progress, but there are certain limitations, such as the limited allelic variation in biparental segregation populations, time consumption for mapping population construction, and limited mapping resolution (Sehgal et al., 2016). In contrast to linkage mapping, GWAS exploits ancestral recombination events in a population, thus providing higher allelic diversity at the loci, resulting in a better association between the marker and the target trait (Kaler et al., 2020).

The application of GWAS to complex quantitative traits of model organisms and crops has increased over the past few years (Atwell et al., 2010; Chen et al., 2014). In soybean, GWAS has successfully identified many high-precision loci associated with yield-related traits. For example, twenty significant SNPs associated with PN have been identified from 211 germplasm by GWAS, and three stable QTL regions were on chromosomes 4, 18 and 20 (Bhat et al., 2022). Wang et al. (2023) used a diverse panel, including 121 wild soybeans, 207 landraces, and 231 improved cultivars to perform GWAS on BM and identified ten important loci, encompassing 47 putative candidate genes. Ayalew et al. (2022) evaluated a germplasm population composed of 541 genotypes and detected 19 QTLs associated with SW by GWAS, of which two stable QTLs on chromosomes 9 and 17 were consistently detected in at least three environments. A large number yield-related loci have been identified, but the genetic basis for production formation regulation has not been fully understood as the complexity of its genetic mechanism, especially under DS conditions.

In this study, we evaluated 188 diverse soybean genotypes under WW and DS conditions across four environments for three yield-related traits, including PN, BM and SW. Furthermore, we used the GWAS approach to analyze genetic loci and key candidate genes related to these traits under WW and DS conditions, which could provide theoretical support for improved yield performance under WW and DS conditions.

2 Materials and methods

2.1 Plant materials and growth conditions

There are 188 diverse genotypes of soybean used in the current GWAS study; which include 95 and 48 genotypes originating from Northeast soybean ecological region and Huanghuaihai region in China, respectively, and 45 genotypes from the United States, Korean, Japan, Russia, etc (Table 1). Of thses, 49 germplasm were landrace, and 139 were improved cultivars. These soybean germplasm were evaluated under WW and DS conditions by both field trials and pot-culture experiments.

Field trials were conducted at Fuxin (121.73788E, 42.13649N) in Liaoning Province, China, in 2018 and 2019 cropping seasons (hereafter referred as FX2018 and FX2019). The climate of this site is a typical semi-arid continental climate with an annual temperature and rainfall of 7.7°C and 450-550 mm, respectively. Three replicates were performed under WW and DS conditions in a randomized block design. Each plot consisted of two rows, 0.6 m apart that were 2 m in length, and the planting density was 165,000 plants per ha. The water supply of WW condition was delivered by drip irrigation, while that of DS treatment was delivered by natural precipitation.

The pot-culture experiments were conducted under open field conditions at Liaoning Academy of Agricultural Sciences, Shenyang (123.56265E, 41.83179N), Liaoning Province, China, in 2020 and 2021 cropping seasons (hereafter referred as SY2020 and SY2021). Soybean seeds were planted in plastic pots (30 cm × 30 cm × 25 cm) with 16.0 kg soil. In a randomized block design, three replications (pots) contained three plants each. The DS treatment was carried out throughout the flowering and podding periods of soybean. Soil moisture content was maintained at 80% of the field's capacity to hold water under WW conditions, whereas it was 60% under water stress conditions. We measured the soil water content every three days and replenished it as needed.

2.2 Phenotypic evaluations and descriptive statistics

Data of three yield-related traits were collected at maturity (R8). In field trials (FX2018 and FX2019), a random sample of 10 plants from each plot were used to determine the yield-related traits, including pod number per plant (PN), biomass per plant (BM) and seed weight per plant (SW). In pot-culture experiments (SY2020 and SY2021), three plants of each pot were used to measure the above traits.

Phenotypic values under WW and DS conditions in the FX2018, FX2019, SY2020 and SY2021 environments were used for analysis. An ANOVA table was used to calculate each trait's broad-sense heritability (Zhao et al., 2020). The best linear unbiased

prediction (BLUP) for each phenotypic value across all environments was calculated using the lmer function in the R package lme4 (http://www.R-project.org/) to reduce environmental variation (Bates et al., 2012). R version 3.5.1 was used to determine Pearson's correlation coefficients (r) for WW and DS conditions separately.

2.3 Genotyping of soybean germplasm

Using a modified CTAB method, DNA from leaves of about 60 d after germination was extracted (Saghai Maroof et al., 1984). SLAF-seq technology (Sun et al., 2013) was used to generate molecular markers in 188 soybean germplasm samples. Our restriction enzymes of choice were *RsaI* and *HaeIII* (NEB, Ipswich, MA, United States) (http://phytozome.jgi.doe.gov/pz/portal.html). Adenine was added to the 3' end of the digested fragments, and the Dual-index was used to distinguish raw sequencing data from digested fragments (Kozich et al., 2013). We obtained SLAF tags by digestion of each soybean germplasm, fragment ligation, PCR amplification, and selection of target fragments for SLAF libraries (Sun et al., 2013). Following quality certification, SLAF-seq using the Illumina HiSeqTM 2500 platform (Illumina, Inc., San Diego, CA, United States) was performed. SLAF libraries were evaluated by comparison them with rice (*Oryza sativa* L. *ssp. japonica* cv. Nipponbare) libraries (http://rice.plantbiology.msu.edu/), which were constructed and sequenced using the same procedures.

In order to ensure the quality of the bioinformatics analysis, a standard protocol was followed in the grouping and genotyping of SLAF-seq data. We compared the filtered sequencing reads with the reference genome using the BWA software (http://bio-bwa.sourceforge.net/) (Li, 2013). In order to classify SLAF makers into polymorphic, non-polymorphic, and repetitive categories, allele frequencies and gene sequence differences were taken into account. SLAF tags were used to identify polymorphic SNP loci mostly using GATK (McKenna et al., 2010). In addition, to ensure the reliability of SNPs identified using GATK, SAMtools also was used to detect SNPs with reference to Li et al. (2009). SNPs that are reliable for further analysis have been identified by both GATK and SAMtools. SNPs with minor allele frequencies (MAF) > 0.05 and marker integrity frequencies > 80% (Zhou et al., 2017) were selected for further analysis.

2.4 Population structure, clustering and linkage disequilibrium analysis

Admixture software was used to generate admixture ratios for K values 1-10 by analyzing population structure 1000 times. Using the valley value of cross-validation error rates, the optimal number of subgroups was determined according to cluster results (Fu and Perry, 2020). Taxonomic and evolutionary relationships between 188 genotypes were assessed using 67,929 SNP markers through phylogenetic analyses. On the basis of the distance matrix, the distance between the materials was calculated using SNP markers from the population. The phylogenetic tree was then constructed using Tree Best (v1.9.2) using the neighbor-joining (NJ) method (Vilella et al., 2009). PopLDdecay software (Zhang et al., 2019) was used to analyze LD for SNPs within a 1 Mb window.

2.5 Genome-wide association studies

A general linear model (GLM) was used for each SNP and trait to test for association between them using TASSEL 5.0. The GLM is based on P + Q matrices, where P is the phenotype matrix and Q is the population structure matrix. The statistical model for the GLM is: y = Xb + e. In this case, y is the data of individual environment or adjusted BLUPs for each trait, X is the known design matrix, b is the fixed effects vector, and e is the random residues vector. A 1000-permutation test was run for the GLM analyses. The Bonferroni-corrected threshold for the p-value was 0.05/67 929 ($p=\alpha/n$, $\alpha=0.05$). For simplicity, $p<7.36E-07$ was used as the threshold value. Manhattan plots were used to visualize significant markers, and quantile-quantile (Q-Q) plots to show important p-value distributions (expected versus observed p-values on a -log10).

2.6 Candidate gene analysis

Based on the GWAS results, pairwise linkage disequilibrium measures were calculated between SNPs in the genomic regions containing significant SNPs. A QTL interval was defined as one where the squared allele frequency correlation between markers was higher than 0.4. We scanned the genome regions in Soybase (www.soybase.org) to identify genes underlying QTLs of interest.

3 Results

3.1 Phenotypic traits evaluation

Three yield-related traits of 188 diverse soybean germplasm was determined under WW and DS conditions in four environments (FX2018, FX2019, SY2020 and SY2021) and the BLUP data for these traits was calculated. The PN, BM and SW under WW and DS conditions exhibited normal distribution, which was basically the same in the four environments as well as the BLUP data (Figure 1). Under WW and DS conditions, as expected, there was significant positive correlations among these yield-related traits. Table 2 shows that PN, BM, and SW had extensive phenotypic variation in soybean germplasm across all four environments. By using BLUP data, the variation ranges of PN, BM and SW under WW condition (hereafter referred as PN-WW, BM-WW and SW-WW) were 21.12-134.52, 22.90-93.08 g, and 2.67-41.08 g, respectively, while those under DS condition (hereafter referred as PN-DS, BM-DS and SW-DS) were 8.66-92.89, 10.93-81.97 g, and 1.18-31.00 g, respectively. The analysis of variance revealed highly significant differences in genotype, environment, and genotype-environment interactions for three yield-related traits. Apart from SW-DS, the effect of environment was larger than that of genotype for these traits. It appears that these yield-related traits are quantitative traits controlled by multiple genes

10.3389/fpls.2023.1265574

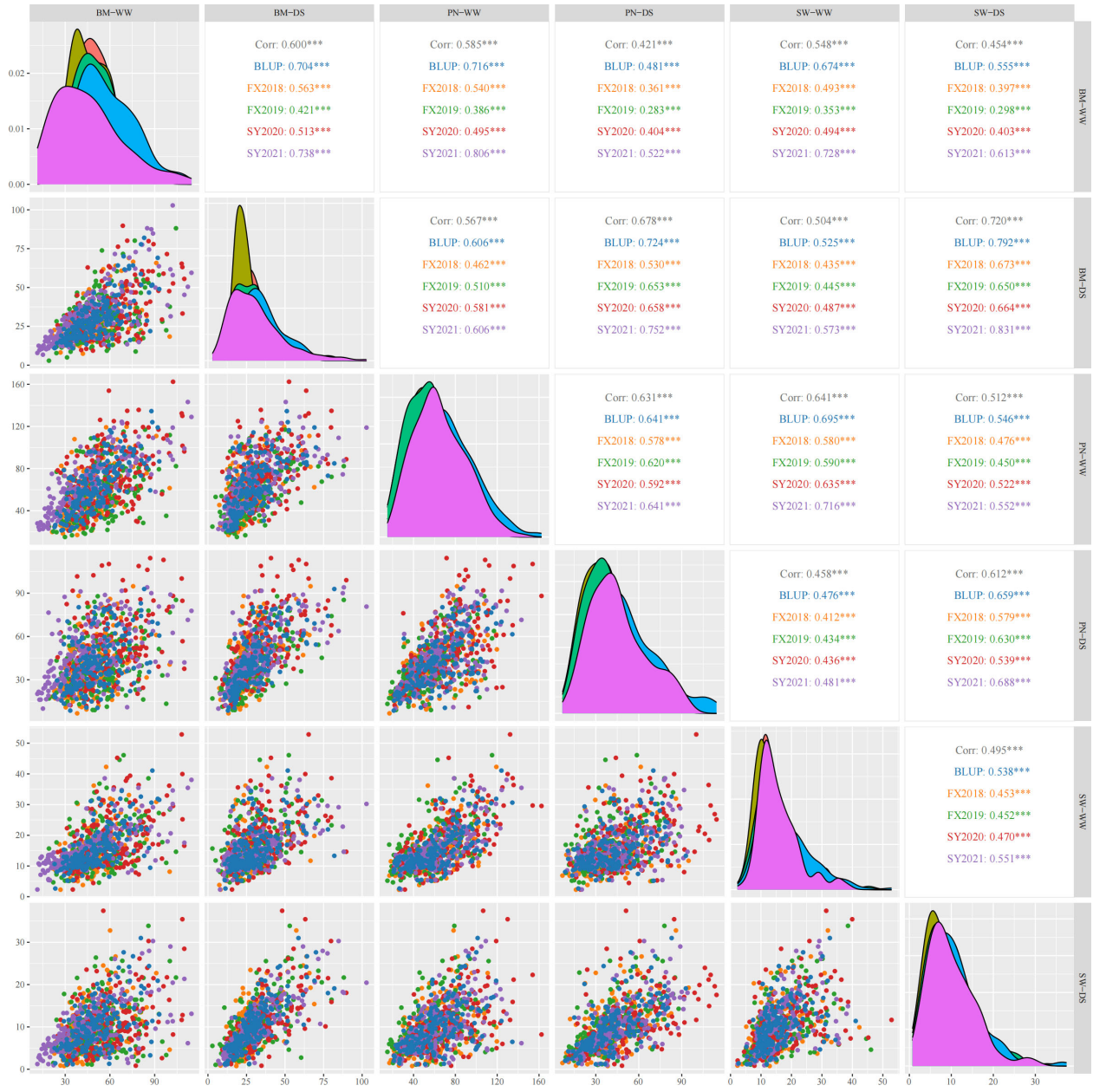


FIGURE 1
Pearson's correlation coefficients describing associations of three yield-related traits evaluated under well-watered (WW) and drought-stressed (DS) conditions in four environments and best linear unbiased prediction (BLUP) data. PN, pod number per plant; BM, biomass per plant; SW, seed weight per plant. The diagonal line illustrates the distribution of six trait-treatments. The scatter plot is displayed below the diagonal line. Above the diagonal line are the correlation coefficient and significant deference. *** represents significant difference at p<0.001.

and easily influenced by environment. The heritability of PN, BM and SW under WW condition was 88%, 86%, and 76%, respectively, while that under DS condition was 88%, 95%, and 85%, respectively.

3.2 Population structure and linkage disequilibrium

Seven subgroups were identified based on the cross-validation error rate and K-values for the 188 genotypes in the Admixture analysis (Figures 2A, B). Further analysis of genetic differentiation was conducted using NJ-based clustering for samples from Northeast and Huang-Huai-Hai regions in China as well as other countries (Figure 2C). According to the phylogenetic tree, there are seven main clusters; each of these groups corresponded to a major subgroup of the Admixture analysis, which supports dividing the population into seven major groups. Further marker-trait association mapping was performed using the Q matrix at K=7. In addition, 188 soybean accessions were assessed for genome-wide LD using a subset of high-quality markers. At a threshold of $r^2 = 0.3$, the average decay distance of LD was 178.7 kb for all 188 soybean accessions (Figure 2D).

10.3389/fpls.2023.1265574

TABLE 1 Geographical source of 188 soybean germplasm in this study.

Geographical source		Landrace	Improved cultivar	Total
Northeast, China	Heilongjiang	9	21	30
	Jilin	13	15	28
	Liaoning	9	25	34
	InnerMongolia	1	5	6
Huang-Huai-Hai, China	Beijing	0	8	8
	Hebei	8	6	14
	Shandong	2	3	5
	Shanxi	3	4	7
	Henan	1	3	4
	Anhui	0	1	1
	Jiangsu	3	3	6
Other country	Korea	0	2	2
	Japan	0	3	3
	Russia	0	2	2
	France	0	2	2
	Italy	0	1	1
	Switzerland	0	1	1
	Ukraine	0	1	1
	US	0	33	33
Total		49	139	188

3.3 GWAS identified significant SNPs associated with yield-related traits

Using a threshold of 7.36E-07, 122 SNPs were significantly associated with PN-WW, BM-WW, SW-WW, PN-DS, BM-DS and SW-DS in the individual environment, which included 40 SNPs in FX2018, 13 SNPs in FX2019, 41 SNPs in SY2020, and 28 SNPs in SY2021 (Supplementary Table S1). By using the BLUP data, a total of 41 SNPs were significantly associated with these traits, as evidenced by the Manhattan and quantile-quantile plots (Q-Q) (Figure 3). For the PN, six significant SNP loci were detected on chromosome 4 and 19 under WW condition, and 12 significant SNP loci were detected on chromosome 8 under DS condition (Figure 3A), which explained about 11-18% of the phenotypic variation (Supplementary Table S1). For the BM, eight significant SNP loci were detected on chromosome 1, 3, 8 and 15 under WW condition, and seven significant SNP loci were detected on chromosome 17 and 18 under DS condition (Figure 3B), which explained about 11-16% of the phenotypic variation (Supplementary Table S1). For the SW, five significant SNP loci were detected on chromosome 1, 4 and 20 under WW condition, and three significant SNP loci were detected on chromosome 8 under DS condition (Figure 3C), which explained about 13-17% of the phenotypic variation (Supplementary Table S1).

3.4 Haplotype analysis in landraces and improved cultivars

In total, 39 significant SNPs were detected simultaneously in the BLUP model and in at least one environment (Supplementary Table S1), which were further used to limit QTL intervals related to the target trait. In the genomic regions of these significant SNPs, the LD blocks were determined. Only 26 QTLs were identified for all 39 significant SNPs, distributed on chromosomes 1, 3, 4, 8, 15, 17, 18, 19, and 20 (Table 2). Of these, six QTL qPN-WW19.1, qPN-DS8.8, qBM-WW1, qBM-DS17.4, qSW-WW4 and qSW-DS8 had at least three significant SNP loci with significant genetic correlation and close genetic relationship. During subsequent haplotype analysis, two or three distinct haplotypes for each QTL were revealed.

QTL qPN-WW19.1 and qPN-DS8.8 that controlled the PN under WW and DS conditions, were detected in approximate interval of 245-kb and 495-kb on chromosome 19 and 8, respectively (Figure 4). For qPN-WW19.1, 91% of landraces and 81% of improved cultivars possessed Hap2, which had greater PN than Hap1 under WW condition.

Two QTL qBM-WW1 and qBM-DS17.4 that controlled the BM under WW and DS conditions, were detected in approximate 184-kb interval on chromosomes 1 and 28-kb interval on chromosomes 17, respectively (Figure 5). For qBM-WW1, only 28% of landraces

10.3389/fpls.2023.1265574

TABLE 2 Descriptive statistics and variance parameters estimated for three traits studied on 188 soybean germplasms under well-watered (WW) and drought-stressed (DS) conditions in four environments and BLUP data.

Environment		PN (/plant)		BM (g/plant)		SW (g/plant)	
		WW	DS	WW	DS	WW	DS
FX2018	Mean	57.54	38.29	47.90	25.59	14.04	9.07
	Std	23.14	17.39	15.06	10.90	6.69	5.45
	CV(%)	40.21	45.41	31.44	42.60	47.63	60.06
	Min	16.95	6.74	18.55	6.51	2.28	0.76
	Max	119.72	94.80	113.06	75.18	42.29	32.79
	H^2	0.95	0.98	0.90	0.95	0.98	0.97
FX2019	Mean	56.35	39.32	52.93	29.75	16.23	9.87
	Std	22.65	17.51	17.44	14.59	7.61	5.87
	CV(%)	40.19	44.54	32.95	49.05	46.87	59.47
	Min	15.05	9.03	17.84	2.48	4.57	1.04
	Max	116.91	87.87	109.52	96.05	46.10	33.88
	H^2	0.95	0.96	0.90	0.97	0.95	0.96
SY2020	Mean	70.69	50.68	58.42	33.79	17.68	10.70
	Std	27.81	22.79	19.24	15.97	8.34	6.24
	CV(%)	39.34	44.96	32.94	47.26	47.19	58.32
	Min	21.17	11.09	22.70	8.36	2.34	0.83
	Max	162.39	114.31	119.76	98.71	52.86	37.45
	H^2	0.96	0.96	0.91	0.97	0.97	0.97
SY2021	Mean	65.48	45.64	45.79	31.06	15.71	10.02
	Std	24.73	19.67	23.25	17.53	6.61	5.64
	CV(%)	37.77	43.09	50.78	56.43	42.07	56.27
	Min	22.00	10.03	9.43	6.62	2.34	0.95
	Max	143.20	94.05	127.29	106.95	38.57	30.29
	H^2	0.95	0.95	0.96	0.97	0.95	0.97
BLUP	Mean	62.31	41.48	50.58	29.65	15.43	9.93
	Std	23.71	17.94	14.93	13.16	6.61	5.49
	CV(%)	38.06	43.24	29.52	44.39	42.84	55.33
	Min	21.12	8.66	22.90	10.93	2.67	1.18
	Max	134.52	92.89	93.08	81.97	41.08	31.00
	H^2	0.88	0.88	0.86	0.95	0.76	0.85
F value	G	228.24***	294.50***	104.63***	306.06***	258.93***	338.82***
	E	874.98***	1327.10***	659.85***	914.71***	614.78***	236.91***
	G × E	5.82***	7.91***	16.43***	17.67***	18.05***	13.17***

BLUP, best linear unbiased prediction; PN, pod number per plant; BM, biomass per plant; SW, seed weight per plant; G, genotype; E, environment; G×E genotype×environment; H^2, broad-sense heritability. *** represents significant difference at $p<0.001$.

and 31% of improved cultivars were included Hap2, which had larger BM than Hap1 under WW condition. For qBM-DS17.4, only 6% of landraces and 10% of improved cultivars were included Hap3, which had larger BM than Hap1 and Hap2 under DS condition.

Two QTL qSW-WW4 and qSW-DS8 that controlled the SW under WW and DS conditons, were detected in approximate 212-kb interval on chromosomes 4 and 12-kb interval on chromosomes 8, respectively (Figure 6). For qSW-WW4, 93% of landraces and 96%

10.3389/fpls.2023.1265574

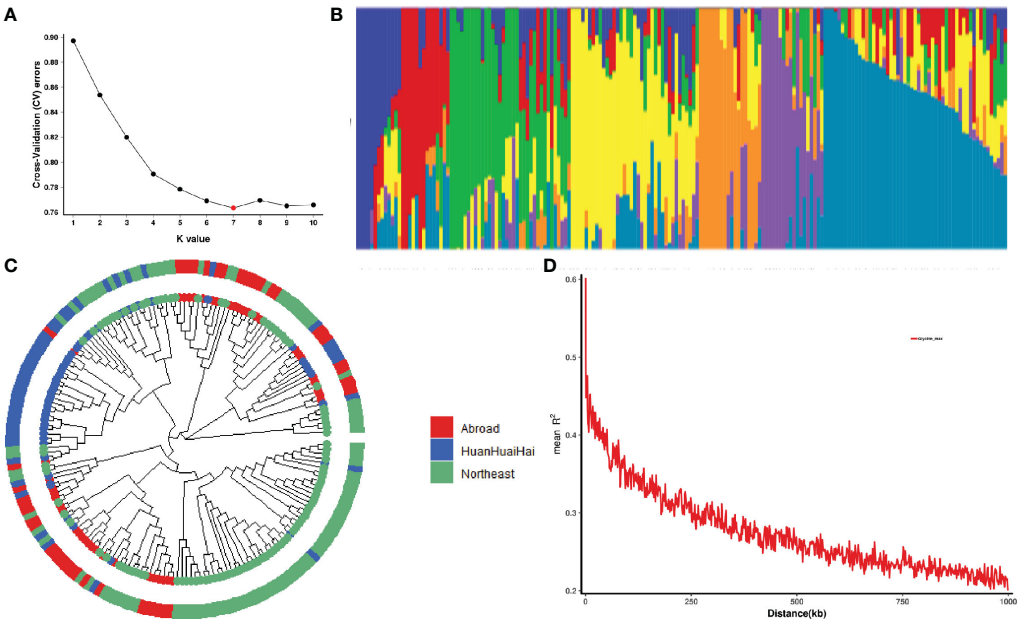


FIGURE 2
Population structure and linkage disequilibrium (LD) analysis of 188 soybean germplasm. **(A)** Cross validation error rate for 188 samples based on clustering from 1 to 10; X-axis is K-value 1-10, Y-axis is cross-validation error rate. **(B)** Colors represent separate groups in clustering analysis when there are seven subgroups. **(C)** Phylogenetic tree of 188 soybean germplasm. Red represents the soybean germplasm from Northeast region, China; Blue represents the soybean germplasm from Huanghuaihai region, China; Green represents the soybean germplasm from other countries. **(D)** A plot of genome-wide LD decay for all 188 soybean germplasm. R^2 indicates the squared allele frequency correlation between each pair of SNP markers. On the X-axis is the distance between each pair of markers.

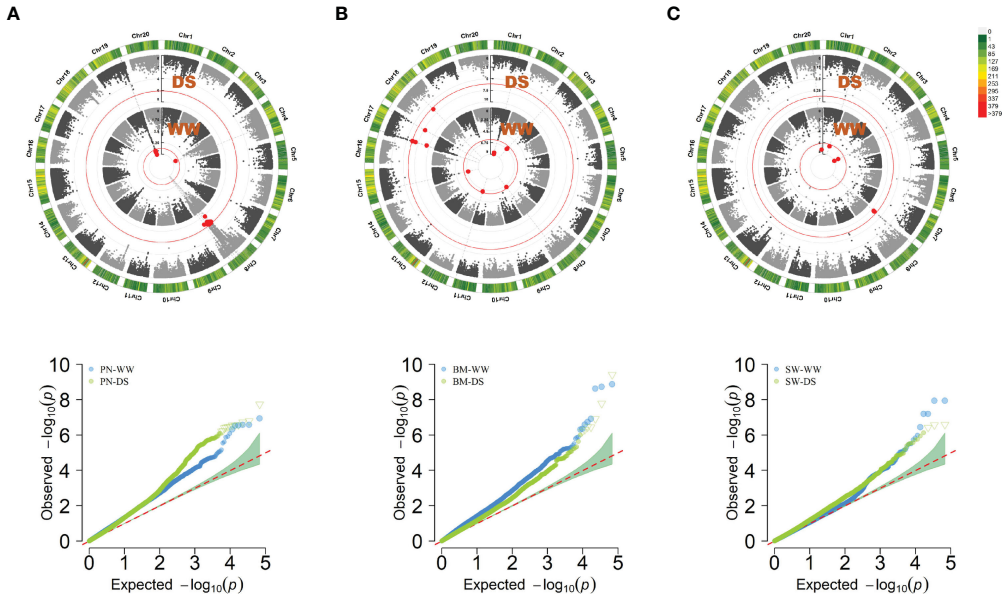


FIGURE 3
Circular manhattan plot and QQ plot for the best linear unbiased prediction (BLUP) values of pod number per plant (PN) **(A)**, biomass per plant (BM) **(B)**, and seed weight per plant (SW) **(C)**, under well-watered (WW) and drought-stressed (DS) conditions, respectively. The *p*-values at the significance thresholds of 7.36E-07.

of improved cultivars were included Hap2 and Hap3, which had higher SW than Hap1 under WW condition. For qSW-DS8, 3% of landraces and 13% of improved cultivars were included Hap2, which had higher SW than Hap1 under DS condition.

3.5 Candidate gene analysis in QTL regions

Using the Glycine max reference genome database (https://www.soybase.org/), we searched for genes associated with yield-

10.3389/fpls.2023.1265574

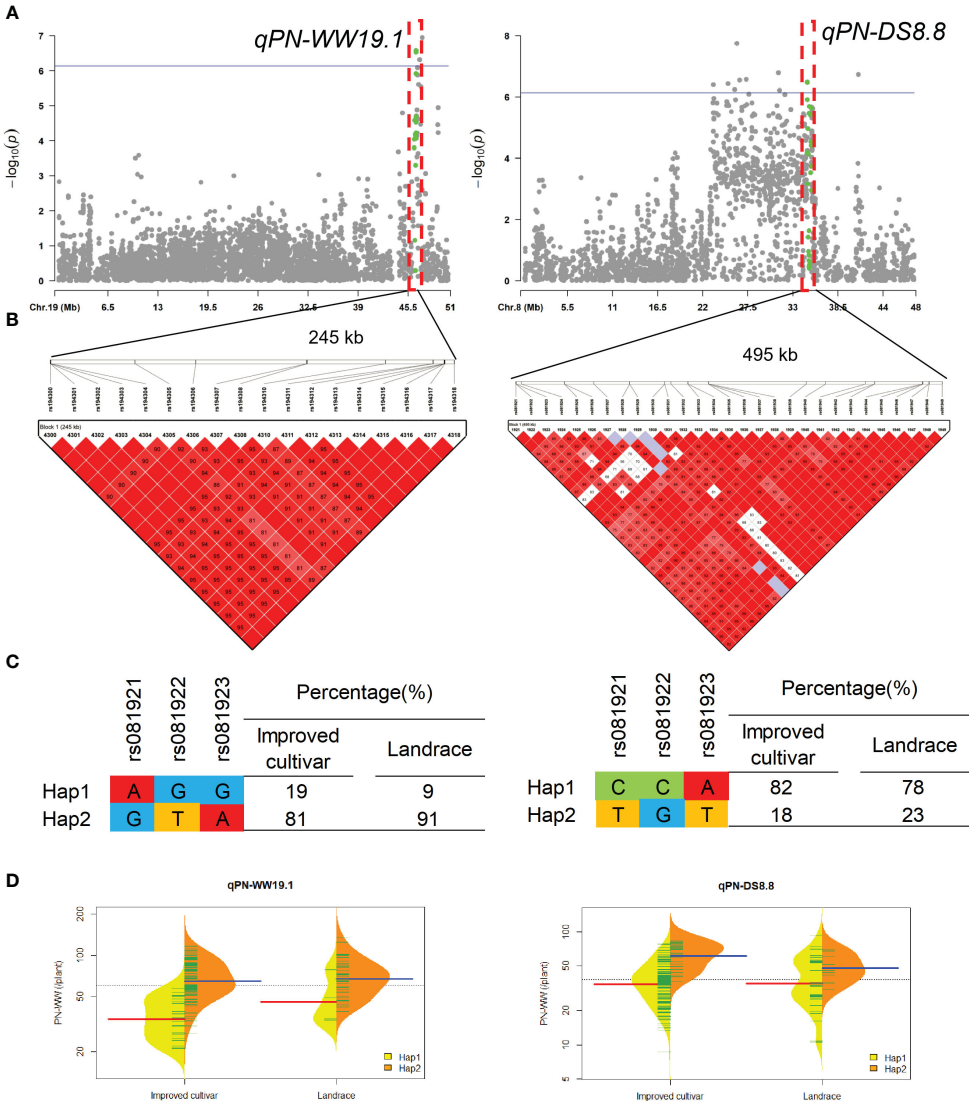


FIGURE 4
Genome-wide association study results for pod number per plant (PN) under well-watered (WW) and drought-stressed (DS) conditions and the analysis of the QTLs qPN-WW19.1 and qPN-DS8.8. **(A)** Manhattan plots for PN under WW and DS conditions. Using the horizontal line as a threshold, the arrows indicate the location of the main peaks. **(B)** Locations of four SNP loci on chromosomes 19 and 8 and their LD based on paired R^2 values. **(C)** 188 soybean germplasm were genotyped by significant SNPs to detect haplotypes. **(D)** Haplotype differences in PN.

related traits and drought tolerance in QTL regions detected under WW and DS conditions, respectively (Table 3). In QTL regions of qSW-WW1, qPN-DS8.3 and qPN-DS8.5, no gene has been found. A total of 208 genes were identified in the 23 remaining QTL regions, and the number of genes varied from 1 to 37 in each QTL region. In this analysis, the number of candidate genes was reduced to 22 genes using annotations based on functional annotations.

Under WW condition, there were three, three, and two candidate genes for PN, BM, and SW, respectively. A total of eight candidate genes were found to be involved in nucleotide transport and metabolism, transcription, carbohydrate transport and metabolism, and cell wall biogenesis. For three important QTL qPN-WW19.1, qBM-WW1 and qSW-WW4, the putative candidate genes were *Glyma.19G211300*, *Glyma.01G119500* and *Glyma.04G124800*, which encoding E3 ubiquitin-protein ligase BAH1, AMP deaminase, and Protein Zinc induced facilitator-like 1, respectively.

In this study, due to their lack of detection under control conditions, the QTLs found under DS conditions were considered drought-responsive. Under DS condition, a total of seven, six and one candidate genes for PN, BM, and SW, respectively, obtained as putative ones for drought responsive in soybean. These 14 candidate genes were involved in transcription, signal transduction mechanisms, secondary metabolites biosynthesis, transport and catabolism, amino acid transport and metabolism, and cell cycle control. For three important QTL qPN-DS8.8, qBM-

10.3389/fpls.2023.1265574

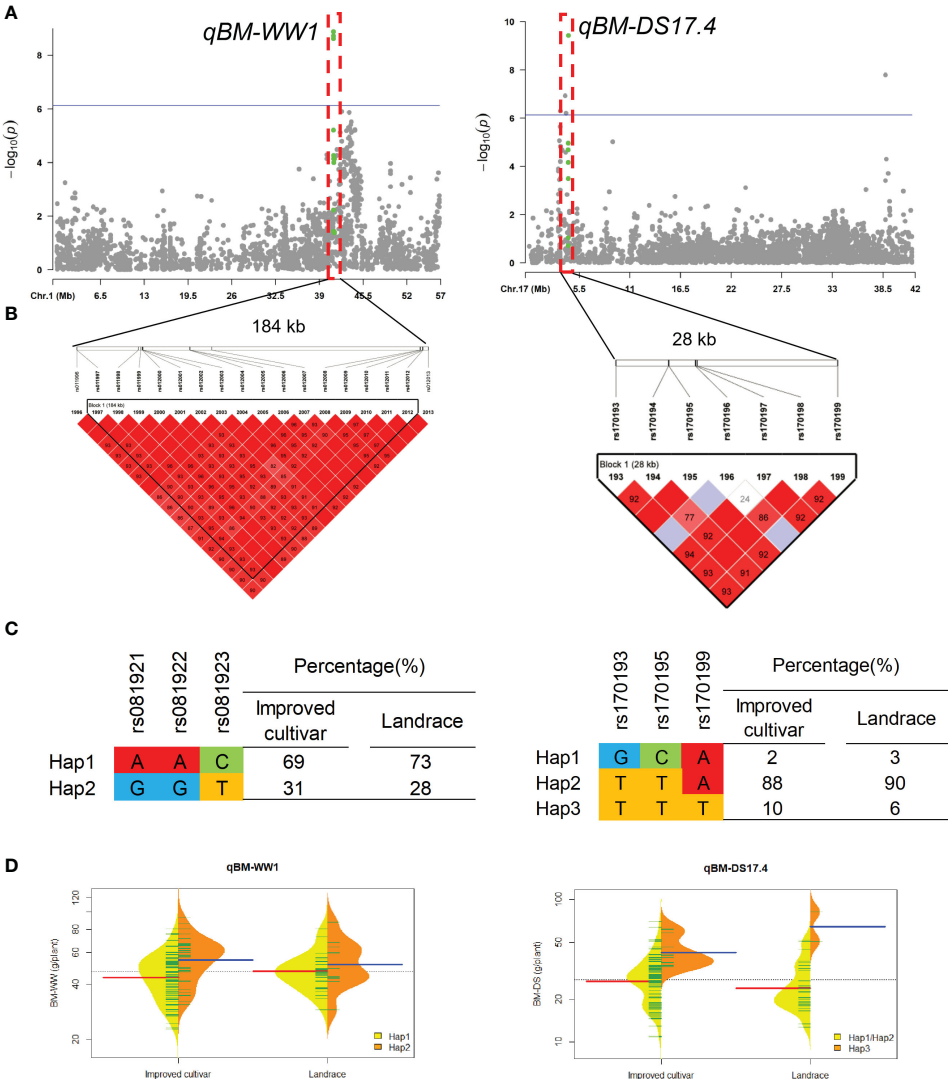


FIGURE 5
Genome-wide association study results for biomass per plant (BM) under well-watered (WW) and drought-stressed (DS) conditions and the analysis of the QTLs qBM-WW1 and qBM-DS17.4. **(A)** Manhattan plots for BM under WW and DS conditions. Using the horizontal line as a threshold, the arrows indicate the location of the main peaks. **(B)** Locations of four SNP loci on chromosomes 1 and 17 and their LD based on paired R^2 values. **(C)** 188 soybean germplasm were genotyped by significant SNPs to detect haplotypes. **(D)** Haplotype differences in BM.

DS17.4 and qSW-DS8, the putative candidate genes were *Glyma.08G269800*, *Glyma.17G057100* and *Glyma.08G020900*, which encoding floral homeotic protein APETALA 1, WRKY transcription factor 11, and ethylene-responsive transcription factor CRF2, respectively.

4 Discussion

Three yield-related traits of 188 soybean germplasm were analyzed under WW and DS conditions in four environments by the GWAS approach. We investigated the genetic basis of phenotypic differences in soybean yield traits, which can serve as a reference for improving soybean molecular breeding under normal as well as drought conditions.

4.1 Yield-related traits analysis

Several complex molecular, physiological, and morphological factors control the reduction in grain yield and yield-related traits under drought stress (Mohammadi, 2014; Kadam et al., 2018). During this experiment, the water deficit was adequate to assess the genotypes' ability to cope with drought, since there was a strong reduction in productivity as well as variations in PN, BM and SW range among accessions. For GWAS analysis, we used BLUP values from four environments to eliminate environmental and locational differences. Both random genetic effects and fixed environments were considered simultaneously in BLUP. It is possible to improve the accuracy of BLUP value prediction by predicting values in different environments and among individuals with different genotypes (Piepho et al., 2008). There has been extensive use of

10.3389/fpls.2023.1265574

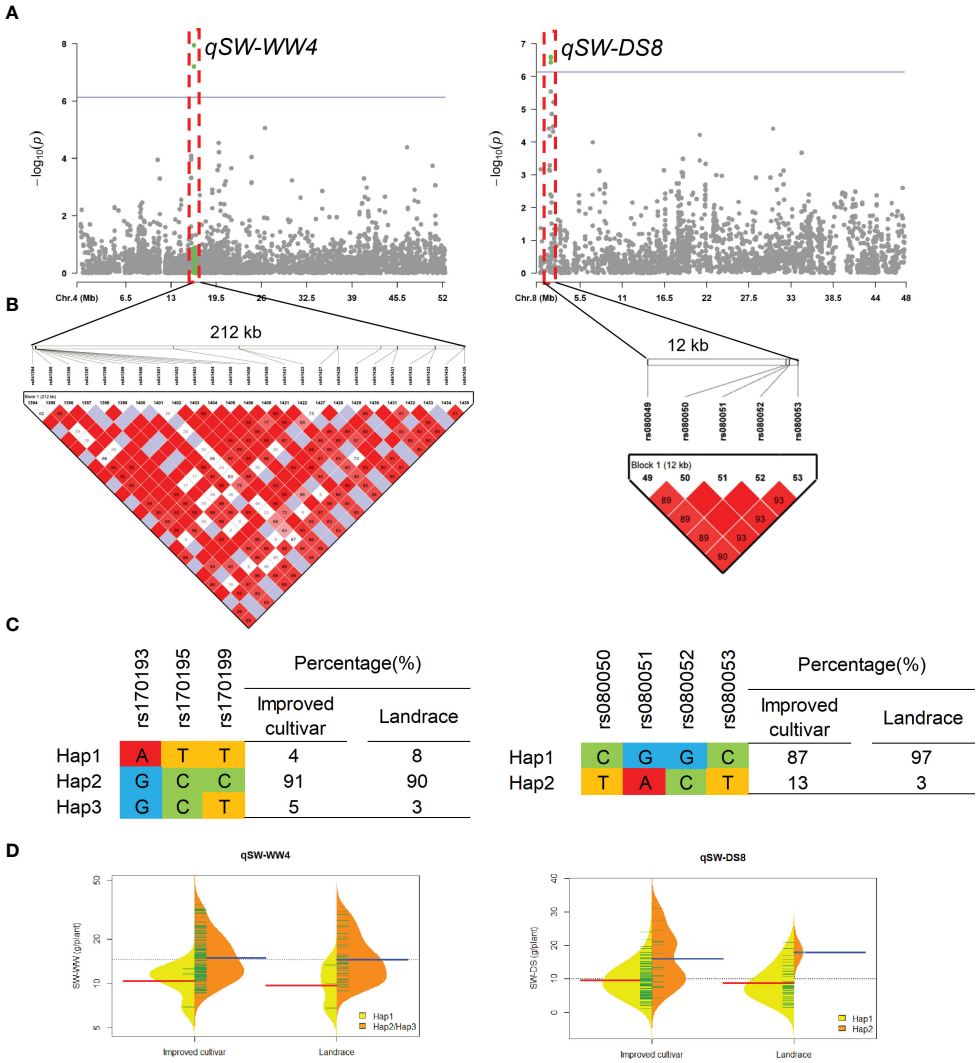


FIGURE 6
Genome-wide association study results for seed weight per plant (SW) under well-watered (WW) and drought-stressed (DS) conditions and the analysis of the QTLs qSW-WW4 and qSW-DS8. **(A)** Manhattan plots for SW under WW and DS conditions. Using the horizontal line as a threshold, the arrows indicate the location of the main peaks. **(B)** Locations of four SNP loci on chromosomes 4 and 8 and their LD based on paired R^2 values. **(C)** 188 soybean germplasm were genotyped by significant SNPs to detect haplotypes. **(D)** Haplotype differences in SW.

this method in QTL mapping, genome-wide association analyses, and the selection of crops based on genome sequences (Wang et al., 2016). Using the BLUP data, large phenotypic variations for the PN, BM and SW can be observed in all the tested materials, especially under DS condition. For all traits scored under WW condition, heritability estimates ranged from 0.76 to 0.88, whereas under DS condition, heritability estimates ranged from 0.85 to 0.95, indicating that these three traits are highly heritable. Therefore, these traits can be used by soybean breeders in selection programs to improve yield and drought tolerance.

4.2 GWAS analysis and gene prediction of key QTLs

By population structure analysis, all the tested materials were divided into seven categories, indicating some variation within the populations. Similar results were found in phylogenetic analyses, suggesting that these analyses can help prevent false positives in GWAS (Eltaher et al., 2018). LD decayed to half the r^2 (0.30) at 178.7 kb, and LD contained a number of significant SNPs, suggesting that GWAS can be used to identify significant markers-trait associations (Schwarz et al., 2015). In the Q-Q diagram analysis results, most points were on the diagonal for all traits, which explains the population structure well (Paterne et al., 2021).

We identified 39 significantly SNPs associated with three traits under WW and DS conditions by BLUP data and individual environmental analyses. For these traits, no overlapping SNPs were observed between WW and DS conditions, which indicates the difficulty of improving soybean yield-related traits simultaneously under different evaluation conditions. Based on the LD analysis, only 26 genomic regions was chosen as the QTL regions with an average of 176-kb intervals.

TABLE 3 List of candidate genes located within the identified QTLs.

Trait	QTL name	Significant SNP	Chr	QTL position	No. of genes	Candidate gene ID	Gene annotation
PN-WW	*qPN-WW4*	rs042851	4	32575507-32592587	0	NA	NA
	qPN-WW19.1	rs194311,rs194316,rs194317	19	46284103-46530081	37	Glyma.19G210900	E3 ubiquitin-protein ligase BAH1
	qPN-WW19.2	rs194339	19	46792316-47006486	25	Glyma.19G217000	WRKY transcription factor 35
	qPN-WW19.3	rs194352	19	47278155-47341447	9	Glyma.19G221600	Polygalacturonase
BM-WW	*qBM-WW1*	rs012000,rs012001,rs012002	1	41066040-41250284	5	Glyma.01G119500	AMP deaminase
	qBM-WW3	rs030672,rs030673	3	6246003-6246085	1	Glyma.03G048500	Disease resistance protein
	qBM-WW15	rs151122	15	17027720-17203787	9	Glyma.15G178700	Eukaryotic translation initiation factor 3
SW-WW	*qSW-WW1*	rs012795	1	51274755		NA	NA
	qSW-WW4	rs041398,rs041399,rs041401	4	16307361-16520021	8	Glyma.04G124800	Protein ZINC INDUCED FACILITATOR-LIKE 1
	qSW-WW20	rs202237	20	37043984-37047052	1	Glyma.20G129100	Protein TIC 21
PN-DS	*qPN-DS8.1*	rs081307	8	23302580-23778598	17	Glyma.08G258800	Aspartic proteinase-like protein 2
	qPN-DS8.2	rs081390	8	24998891-25197286	3	Glyma.08G261200	Homocysteine S-methyltransferase 1
	qPN-DS8.3	rs081420	8	25808374-26012431	0	NA	NA
	qPN-DS8.4	rs081438	8	26079563-26228773	1	Glyma.08G261700	NA
	qPN-DS8.5	rs081461,rs081462	8	26498358-26498365	0	NA	NA
	qPN-DS8.6	rs081509	8	27277729-27586943	1	Glyma.08G262500	U-box domain-containing protein 14
	qPN-DS8.7	rs081715,rs081722	8	30932153-31426811	3	Glyma.08G265200	Calcium-binding protein CML21
	qPN-DS8.8	rs081921,rs081922,rs081923	8	34768849-35264470	12	Glyma.08G269800	Floral homeotic protein APETALA 1
	qPN-DS8.9	rs082209	8	40748092-40990195	18	Glyma.08G293300	Transcription factor MYB1R1
BM-DS	*qBM-DS17.1*	rs170162	17	3412747-3466772	1	Glyma.17G045900	Embryogenesis-associated protein EMB8
	qBM-DS17.2	rs170177	17	3870840-4016954	19	Glyma.17G052200	UBP1-associated proteins 1C
	qBM-DS17.3	rs170185	17	4029241-4068527	7	Glyma.17G053500	Casein kinase 1-like protein 1
	qBM-DS17.4	rs170193,rs170195,rs170199	17	4294571-4322918	20	Glyma.17G057100	WRKY transcription factor 11
	qBM-DS17.5	rs173557	17	38770868-38770949	1	Glyma.17G232500	RNA-binding protein 1

(Continued)

TABLE 3 Continued

Trait	QTL name	Significant SNP	Chr	QTL position	No. of genes	Candidate gene ID	Gene annotation
	qBM-DS18	rs184014,rs184015	18	37970702-38400499	8	Glyma.18G164100	1-aminocyclopropane-1-carboxylate oxidase homolog 12
SW-DS	*qSW-DS8*	rs080050,rs080051,rs080052, rs080053	8	1692570-1704747	2	Glyma.08G020900	Ethylene-responsive transcription factor CRF2

Six QTL regions containing at least three significant SNP loci with significant LD tend to co-inherit, which can be useful for further genetic validation as well as marker-assisted selection. Among these QTLs, three were consistent with previously reported soybean QTLs. For example, within the previous reported QTL interval (Chr19:386234-49312675) controlling PN (Zhang J. et al., 2015), the present QTL qPN-WW19.1 associated PN under WW condition was detected in SY2020, FX2021 and BLUP data. Moreover, one SNP loci (Chr19:46340503) significantly associate with plant height in soybean was previously reported by Fang et al. (2017), which was also located within the interval of qPN-WW19.1 (Chr19: 46284103-46530081). Within the QTL interval of qPN-WW19.1, a gene *Glyma.19G211300*, encoding E3 ubiquitin-protein ligase BAH1, was predicted here as the putative candidate gene. Members of the protein family E3 ubiquitin-protein ligases play a significant role in the ubiquitin-proteasome pathway to affect yield (Ge et al., 2016; Lv et al., 2022), such as GW2 in rice (Choi et al., 2018), ZmGW2 in maize (Kong et al., 2014), and TaGW2 in wheat (Lv et al., 2022).

The QTL qBM-DS17.4 associated BM under DS condition was detected in FX2018, SY2020, FX2021 and BLUP data, which located within the previous reported QTL interval (Chr17:5891979-4629130) controlling shoot dry weight in soybean (Liang et al., 2010). Within the QTL interval of qBM-DS17.4, a gene *Glyma.17G057100*, encoding WRKY transcription factor 11, was predicted here as the putative candidate gene. WRKY transcription factors participate in various physiological and developmental processes (Rushton et al., 2010), such as seed development (Lagacã and Matton, 2004), seed dormancy and germination (Zentella et al., 2007), senescence (Silke and Imre, 2002), and development (Johnson et al., 2002). Plant hormones, including abscisic acid (Zhang L. et al., 2015), jasmonic acid (Shimono et al., 2007) and gibberellin (Zhang L. et al., 2015), are signaled by WRKY proteins, according to recent findings. WRKY transcription factors have been demonstrated to confer drought tolerance in wheat (Gao et al., 2018; El-Esawi et al., 2019) and soybean (Zhou et al., 2008; Shi et al., 2018).

The QTL qSW-WW4 associated SW under WW condition was detected in FX2018, SY2020 and BLUP data, which located within the previous reported QTL interval (Chr17:12310119-32617784) that evaluated for the SW for a population grown in a low phosphorus environment (Liang et al., 2010). Within the QTL interval of qBM-DS17.4, a gene *Glyma.04G124800*, encoding Protein Zinc induced facilitator-like 1, was predicted here as the putative candidate gene. Due to their specialized role in phytosiderophores efflux and auxin homeostasis, a subset of the zinc-induced facilitators are also proven to impart tolerance to micronutrient deficiencies. In the case of Zn deficiency, crop yield is affected (Krithika and Balachandar, 2016), while Fe deficiency can impair several vital functions, such as photosynthesis and respiration (Marschner, 1995). ZIFL genes contributes to mobilization of Zn^{2+} in rhizospheric regions and mobilization of Fe there by secreting phytosiderophores (Haydon and Cobbett, 2007; Meena et al., 2021)

QTL are considered validated if they are detected in a different background as it is a true association across many genotypes. In this study, all QTLs detected except the validated ones can be considered novel locus that should be tested in another population. For example, within the QTL interval of qSW-DS8, a gene *Glyma.08G020900*, encoding ethylene-responsive transcription factor CRF2, was predicted here as the putative candidate gene. In many species, members of the AP2/ERF superfamily regulate flower and seed development, and thus play a critical role in regulating seed weight and further controlling seed yield (Jiang et al., 2020). A subfamily of ERF proteins called cytokinin response factors (CRFs) contributes to plant growth, development, nitrogen uptake, and stress resistance (Zong et al., 2021). Recently, the gene GmCRF4a in soybean has been reported to regulate plant height and auxin biosynthesis, which would facilitate future molecular breeding practice to improve soybean architecture (Xu et al., 2022).

4.3 Favorable haplotypes for soybean breeding

Using the base types of SNP markers and distributions of alleles associated with a trait, some haplotypes were identified, and favorable haplotypes were identified based on their phenotypic values using t-tests. The cultivars with favorable haplotypes in qPN-WW19.1, qBM-WW1 and qSW-WW4 usually had greater PN, BM and SW, respectively, under WW condition, while those in qBM-DS17.4, qPN-DS8.8 and qSW-DS8 also had more desirable phenotypes, respectively, under DS condition. During soybean breeding, these important QTLs had been subjected to various levels of selection, resulting in different proportions of favorable haplotypes for each locus.

It has been well documented that the development of soybean breeding has led to a change in agronomic traits. Linear increases in PN and SW accounted for most of the historical yield improvement (Morrison et al., 2000; Cui and Yu, 2005; Jin et al., 2010). In this study, we found larger proportions of favorable haplotypes for locus qPN-WW19.1 and qSW-WW4 in both landraces and improved

cultivars, suggesting the selection for these favorable haplotypes by breeders played an important role during historical yield improvement. In this study, about 59.04% of the population, including improved cultivar 'Liaodou69' (32.60 g/plant), 'Liaodou32' (31.92 g/plant), 'Liaodou36' (31.54 g/plant), 'Liaodou14' (30.49 g/plant), 'Zhonghuang35' (30.03 g/plant), and 'Tiefeng31' (28.04 g/plant) carried both superior haplotypes for locus qPN-WW19.1 and qSW-WW4 and produced greater yields under WW condition, suggesting that these QTLs had aggregated by soybean breeding. Although the historical yield improvement was primarily driven by higher BM (Balboa et al., 2018), we found less proportions of favorable haplotypes for qBM-WW1, especially in landraces. Moreover, the proportions of favorable haplotypes for locus qBM-DS17.4, qPN-DS8.8 and qSW-DS8 were only 23%, 6% and 3% in landraces, respectively, even though in improved cultivars those were 18%, 10% and 13%, respectively. It may be due to the belief that crop improvement has reduced their ability to cope with future challenges, such as drought (Byrne et al., 2018; Swarup et al., 2020). Our results implied that these QTLs qBM-DS17.4, qPN-DS8.8 and qSW-DS8 had not experienced strong selection during drought tolerant soybean breeding but had potential for increasing soybean drought tolerance.

5 Conclusion

In this study, we genotyped 188 soybean germplasm using SLAF-seq technology and evaluated their yield-related traits under WW and DS conditions. By using BLUP data and individual environmental analyses in GWAS, a total of 39 SNPs were significantly associated with three traits under two conditions, which were tagged to 26 genomic regions by linkage disequilibrium (LD) analysis. Six locus could play a key role in determining PN, BM and SW of soybean. The favorable haplotypes for locus qPN-WW19.1 and qSW-WW4 had experienced strong selection during historical yield improvement, while those for qBM-WW1, qBM-DS17.4, qPN-DS8.8 and qSW-DS8 had not been fully utilized, especially for drought tolerant soybean breeding. It was believed that the superior haplotypes for these loci should be integrated to improve yield-related traits. As a result of this study, a better understanding of the genetic architecture driving high yields will be gained and the foundation for marker-assisted breeding will be laid in soybean.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: NCBI SRA database accession number PRJNA1014913.

Author contributions

SL: Data curation, Funding acquisition, Investigation, Project administration, Writing- original draft, Writing- review & editing. YC: Investigation, Formal analysis, Writing- review & editing. CW: Investigation, Methodology, Writing- review & editing. CY: Investigation, Data curation, Writing- review & editing. XS: Investigation, Methodology, Writing- review & editing. LZ: Data curation, Investigation, Writing- review & editing. WW: Project administration, Writing- review & editing. SS: Project administration, Writing- review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. The study was supported by the National Natural Science Foundation of China (32101795 and 32301782), Liaoning provincial Major Special Project of Agricultural Science and Technology (2022JH1/10200002 and 2021JH1/10400038), Key Research and Development Plan of Liaoning Science and Technology Department (2021JH2/1020027), and Shenyang Seed Industry Innovation Project (22-318-2-12).

Acknowledgments

We also acknowledge Dr. Mingzhu Zhao, Institute of Crop Research, Liaoning Academy of Agricultural Sciences, for comments improving the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1265574/full#supplementary-material

References

Atwell, S., Huang, Y. S., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Ayalew, H., Schapaugh, W., Vuong, T., and Nguyen, H. T. (2022). Genome-wide association analysis identified consistent QTL for seed yield in a soybean diversity panel tested across multiple environments. *Plant Genome* 15, e20268. doi: 10.1002/tpg2.20268

Bates, D., Maechler, M., and Bolker, B. (2012) *lme4: linear mixed-effects models using S4 classes (R package version 0.999999-0)*. Available at: http://cran.r-project.org/web/packages/lme4/index.html.

Bhat, J. A., Adeboye, K. A., Ganie, S. A., Barmukh, R., Hu, D., Varshney, R. K., et al. (2022). Genome-wide association study, haplotype analysis, and genomic prediction reveal the genetic basis of yield-related traits in soybean (Glycine max L.). *Front. Genet.* 13. doi: 10.3389/fgene.2022.953833

Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., and Smith, S. (2018). Sustaining the future of plant breeding: the critical role of the USDA-ARS National Plant Germplasm System. *Crop Sci.* 58, 451–468. doi: 10.2135/cropsci2017.05.0303

Cerezini, P., Kuwano, B. H., dos Santos, M. B., Terassi, F., Hungria, M., and Nogueira, M. A. (2016). Strategies to promote early nodulation in soybean under drought. *Field Crops Res.* 196, 160–167. doi: 10.1016/j.fcr.2016.06.017

Chen, W., Gao, Y. Q., Xie, W. B., Gong, L., Lu, K., Wang, W. S., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Chen, W., Yao, Q., Patil, G. B., Agarwal, G., Deshmukh, R. K., Lin, L., et al. (2016). Identification and comparative analysis of differential gene expression in soybean leaf tissue under drought and flooding stress revealed by RNA-Seq. *Front. Plant Sci.* 7, 1044. doi: 10.3389/fpls.2016.01044

Choi, B. S., Kim, Y. J., Markkandan, K., Koo, Y. J., Song, J. T., and Seo, H. S. (2018). GW2 functions as an E3 ubiquitin ligase for rice expansin-like 1. *Int. J. Mol. Sci.* 19 (7), 1904. doi: 10.3390/ijms19071904

Cui, S. Y., and Yu, D. Y. (2005). Estimates of relative contribution of biomass, harvest index and yield components to soybean yield improvements in China. *Plant Breed.* 124, 473–476. doi: 10.1111/j.1439-0523.2005.01112.x

El-Esawi, M. A., Al-Ghamdi, A. A., Ali, H. M., and Ahmad, M. (2019). Overexpression of AtWRKY30 transcription factor enhances heat and drought stress tolerance in wheat (Triticum aestivum L.). *Genes* 10, 163. doi: 10.3390/genes10020163

Eltaher, S., Sallam, A., Belamkar, V., Emara, H. A., Nower, A. A., Salem, K. F., et al. (2018). Genetic diversity and population structure of F3: 6 Nebraska winter wheat genotypes using genotyping-by-sequencing. *Front. Genet.* 9, 76. doi: 10.3389/fgene.2018.00076

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18, 161. doi: 10.1186/s13059-017-1289-9

Frederick, J. R., Camp, C. R., and Bauer, P. J. (2001). Drought-stress effects on branch and mainstem seed yield and yield components of determinate soybean. *Crop Sci.* 41, 759–763. doi: 10.2135/cropsci2001.413759x

Fu, W., and Perry, P. O. (2020). Estimating the number of clusters using cross-validation. *J. Comput. Graph. Stat.* 29, 162–173. doi: 10.1080/10618600.2019.1647846

Gao, H., Wang, Y., Xu, P., and Zhang, Z. (2018). Overexpression of a WRKY transcription factor TaWRKY2 enhances drought stress tolerance in transgenic wheat. *Front. Plant Sci.* 9, 997. doi: 10.3389/fpls.2018.00997

Ge, L., Yu, J., Wang, H., Luth, D., Bai, G., Wang, K., et al. (2016). Increasing seed size and quality by manipulating BIG SEEDS1 in legume species. *Proc. Natl. Acad. Sci.* 113 (44), 12414–12419. doi: 10.1073/pnas.1611763113

Hacisalihoglu, G., Burton, A. L., Gustin, J. L., Eker, S., Asikli, S., Heybet, E. H., et al. (2018). Quantitative trait loci associated with soybean seed weight and composition under different phosphorus levels. *J. Integr. Plant Biol.* 60, 232–241. doi: 10.1111/jipb.12612

Haydon, M. J., and Cobbett, C. S. (2007). A novel major facilitator superfamily protein at the tonoplast influences zinc tolerance and accumulation in Arabidopsis. *Plant Physiol.* 143, 1705–1719. doi: 10.1104/pp.106.092015

Jiang, W., Zhang, X., Song, X., Yang, J., and Pang, Y. (2020). Genome-wide identification and characterization of APETALA2/ethylene-responsive element binding factor superfamily genes in soybean seed development. *Front. Plant Sci.* 11, 566647. doi: 10.3389/fpls.2020.566647

Jin, J., Liu, X., Wang, G., Mi, L., Shen, Z., Chen, X., et al. (2010). Agronomic and physiological contributions to the yield improvement of soybean cultivars released from 1950 to 2006 in Northeast China. *Field Crops Res.* 115, 116–123. doi: 10.1016/j.fcr.2009.10.016

Johnson, C. S., Kolevski, B., and Smyth, D. R. (2002). Transparent TESTA glabra2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. *Plant Cell* 14, 1359–1375. doi: 10.1105/tpc.001404

Kadam, N. N., Struik, P. C., Rebolledo, M. C., Yin, X., and Jagadish, S. K. (2018). Genome-wide association reveals novel genomic loci controlling rice grain yield and its component traits under water-deficit stress during the reproductive stage. *J. Exp. Bot.* 69, 4017–4032. doi: 10.1093/jxb/ery186

Kaler, A. S., Gillman, J. D., Beissinger, T., and Purcell, L. C. (2020). Comparing different statistical models and multiple testing corrections for association mapping in soybean and maize. *Front. Plant Sci.* 10, 1794. doi: 10.3389/fpls.2019.01794

Kong, M., Li, C., Sun, Q., Lu, M., Wang, W., Pan, J., et al. (2014). Isolation and expression analysis of the E3 ubiquitin ligase encoding gene ZmGW2-1 in maize. *J. Anhui Agric. Univ.* 41 (6), 1004–1011. doi: 10.13610/j.cnki.1672-352x.20141029.009

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., and Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* 79, 5112–5120. doi: 10.1128/AEM.01043-13

Krithika, S., and Balachandar, D. (2016). Expression of zinc transporter genes in rice as influenced by zinc-solubilizing enterobacter cloacae strain ZSB14. *Front. Plant Sci.* 7, 446. doi: 10.3389/fpls.2016.00446

Lagacã, M., and Matton, D. P. (2004). Characterization of a WRKY transcription factor expressed in late torpedo-stage embryos of Solanum chacoense. *Planta* 219, 185–189. doi: 10.1007/s00425-004-1253-2

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv*, 1–3. doi: 10.48550/arXiv.1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Liang, Q., Cheng, X., Mei, M., Yan, X., and Liao, H. (2010). QTL analysis of root traits as related to phosphorus efficiency in soybean. *Ann. Bot.* 106, 223–234. doi: 10.1093/aob/mcq097

Liu, F., Jensen, C. R., and Andersen, M. N. (2004). Drought stress effect on carbohydrate concentration in soybean leaves and pods during early reproductive development: its implication in altering pod set. *Field Crops Res.* 86, 1–13. doi: 10.1016/S0378-4290(03)00165-5

Lv, Q., Li, L., Meng, Y., Sun, H., Chen, L., Wang, B., et al. (2022). Wheat E3 ubiquitin ligase TaGW2-6A degrades TaAGPS to affect seed size. *Plant Sci.* 320, 111274. doi: 10.1016/j.plantsci.2022.111274

Marschner, H. (1995). "9—Functions of mineral nutrients: micronutrients," in *Mineral nutrition of higher plants*, *2nd ed*. Ed. H. Marschner (London, UK: Academic Press), 313–404.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Meena, V., Sharma, S., Kaur, G., Singh, B., and Pandey, A. K. (2021). Diverse functions of plant zinc-induced facilitator-like transporter for their emerging roles in crop trait enhancement. *Plants* 11, 102. doi: 10.3390/plants11010102

Mohammadi, R., Haghparast, R., Sadeghzadeh, B., Ahmadi, H., Solimani, K., and Amri, A. (2014). Adaptation patterns and yield stability of durum wheat landraces to highland cold rainfed areas of Iran. *Crop Sci.* 54, 944–954. doi: 10.2135/cropsci2013.05.0343

Morrison, M. J., Voldeng, H. D., and Cober, E. R. (2000). Agronomic changes from 58 years of genetic improvement of short-season soybean cultivars in Canada. *Agron. J.* 92, 780–784. doi: 10.2134/agronj2000.924780x

Paterne, A. A., Norman, P. E., Asiedu, R., and Asfaw, A. (2021). Identification of quantitative trait nucleotides and candidate genes for tuber yield and mosaic virus tolerance in an elite population of white Guinea yam (Dioscorea rotundata) using genome-wide association scan. *BMC Plant Biol.* 21, 552. doi: 10.1186/s12870-021-03314-w

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

Rushton, P. J., Somssich, I. E., Ringler, P., and Shen, Q. J. (2010). WRKY transcription factors. *Plant Signal. Behav.* 15, 247–258. doi: 10.1016/j.tplants.2010.02.006

Saghai Maroof, M. A., Soliman, K. M., Jorgensen, R. A., and Allard, R. W. (1984). Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. *Proc. Natl. Acad. Sci. U. S. A.* 81, 8014–8018. doi: 10.1073/pnas.81.24.8014

Schwarz, N., Armbruster, U., Iven, T., Brückle, L., Melzer, M., Feussner, I., et al. (2015). Tissue-specific accumulation and regulation of zeaxanthin epoxidase in arabidopsis reflect the multiple functions of the enzyme in plastids. *Plant Cell Physiol.* 56, 346–357. doi: 10.1093/pcp/pcu167

Sehgal, D., Singh, R., and Rajpal, V. R. (2016). "Quantitative trait loci mapping in plants: Concepts and approaches," in *Molecular breeding for sustainable crop improvement*, vol. 2 . Eds. V. R. Rajpal, S. R. Rao and S. N. Raina (Cham, Switzerland: Springer International Publishing), 31–59.

Shi, W. Y., Du, Y. T., Ma, J., Min, D. H., Jin, L. G., Chen, J., et al. (2018). The WRKY transcription factor GmWRKY12 confers drought and salt tolerance in soybean. *Int. J. Mol. Sci.* 19, 4087. doi: 10.3390/ijms19124087

Shimono, M., Sugano, S., Nakayama, A., Jiang, C. J., Ono, K., Toki, S., et al. (2007). Rice WRKY45 plays a crucial role in benzothiadiazole-inducible blast resistance. *Plant Cell* 19, 2064–2076. doi: 10.1105/tpc.106.046250

Silke, R., and Imre, E. S. (2002). Targets of AtWRKY6 regulation during plant senescence and pathogen defense. *Genes Dev.* 16, 1139–1149. doi: 10.1101/gad.222702

Sun, X. W., Liu, D. Y., Zhang, X. F., Li, W. B., Liu, H., Hong, W. G., et al. (2013). SLAF-seq: An efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. *PLoS One* 8, e58700. doi: 10.1371/journal.pone.0058700

Sun, X., Sun, X., Pan, X., Zhang, H., Wang, Y., Ren, H., et al. (2022). Identification and fine mapping of a quantitative trait locus controlling the total flower and pod numbers in soybean. *Agronomy* 12, 790. doi: 10.3390/agronomy12040790

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., and glenn, K. C. (2020). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* 61, 839–852. doi: 10.1002/csc2.20377

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). Ensem-blCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. doi: 10.1101/gr.073585.107

Wang, C. L., Zhang, Q., Jiang, L., Qian, R., Ding, X. D., and Zhao, Y. F. (2016). Comparative study of estimation methods for genomic breeding values. *Sci. Bull.* 61, 353–356. doi: 10.1007/s11434-016-1014-1

Wang, X., Zhou, S., Wang, J., Lin, W., Yao, X., Su, J., et al. (2023). Genome-wide association study for biomass accumulation traits in soybean. *Mol. Breed.* 43, 33. doi: 10.1007/s11032-023-01380-6

Xu, Z., Wang, R., Kong, K., Begum, N., Almakas, A., Liu, J., et al. (2022). An APETALA2/ethylene responsive factor transcription factor GmCRF4a regulates plant height and auxin biosynthesis in soybean. *Front. Plant Sci.* 13, 983650. doi: 10.3389/fpls.2022.983650

Yang, Y. (2021). *Fine mapping and candidate gene identification of a soybean seed protein and oil qtl from a wild soybean accession and linkage analysis for whole plant biomass, carbon, nitrogen, and seed composition using a RIL mapping population* (Columbia, MO, United States: Doctoral dissertation, University of Missouri-Columbia).

Yoosefzadeh Najafabadi, M. (2021). *Using advanced proximal sensing and genotyping tools combined with bigdata analysis methods to improve soybean yield* (Guelph, ON, Canada: University of Guelph).

Zentella, R., Zhang, Z., Park, M., Thomas, S., Endo, A., Murase, K., et al. (2007). Global analysis of della direct targets in early gibberellin signaling in Arabidopsis. *Plant Cell* 19, 3037–3057. doi: 10.1105/tpc.107.054999

Zhang, C., Dong, S. S., Xu, J. Y., He, W. M., and Yang, T. L. (2019). PopLDdecay: A fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* 35, 1786–1788. doi: 10.1093/bioinformatics/bty875

Zhang, L., Gu, L., Ringler, P., Smith, S., Rushton, P. J., and Shen, Q. J. (2015). Three WRKY transcription factors additively repress abscisic acid and gibberellin signaling in aleurone cells. *Int. J. Exp. Plant Biol.* 236, 214–222. doi: 10.1016/j.plantsci.2015.04.014

Zhang, J., Song, Q., Cregan, P. B., Nelson, R. L., Wang, X., Wu, J., et al. (2015). Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (Glycine max) germplasm. *BMC Genomics* 16, 1–11. doi: 10.1186/s12864-015-1441-4

Zhao, M., Ma, Z., Wang, L., Tang, Z., Mao, T., Liang, C., et al. (2020). SNP-based QTL mapping for panicle traits in the japonica super rice cultivar Liaoxing 1. *Crop J.* 8, 769–780. doi: 10.1016/j.cj.2020.07.002

Zhou, Q. Y., Tian, A. G., Zou, H. F., Xie, Z. M., Lei, G., Huang, J., et al. (2008). Soybean WRKY-type transcription factor genes, GmWRKY13, GmWRKY21, and GmWRKY54, confer differential tolerance to abiotic stresses in transgenic Arabidopsis plants. *Plant Biotechnol. J.* 6, 486–503. doi: 10.1111/j.1467-7652.2008.00336.x

Zhou, Q., Zhou, C., Zheng, W., Mason, A. S., Fan, S., Wu, C., et al. (2017). Genome-Wide SNP markers based on SLAF-Seq uncover breeding traces in rapeseed (Brassica napus L.). *Front. Plant Sci.* 8, 648. doi: 10.3389/fpls.2017.00648

Zong, Y., Hao, Z., Tu, Z., Shen, Y., Zhang, C., Wen, S., et al. (2021). Genome-wide survey and identification of AP2/ERF genes involved in shoot and leaf development in Liriodendron chinense. *BMC Genomics* 22, 807. doi: 10.1186/s12864-021-08119-7

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 October 2023
DOI 10.3389/fpls.2023.1153208

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Zhansheng Li,
Chinese Academy of Agricultural Sciences, China
Qiusheng Kong,
Huazhong Agricultural University, China
Jianbin Hu,
Henan Agricultural University, China

*CORRESPONDENCE
Jiaowen Cheng
jiaolong1015@126.com
Kailin Hu
hukailin@scau.edu.cn

RECEIVED 29 January 2023
ACCEPTED 25 September 2023
PUBLISHED 10 October 2023

CITATION
Zhong J, Cui J, Miao M, Hu F, Dong J, Liu J, Zhong C, Cheng J and Hu K (2023) A point mutation in *MC06g1112* encoding FLOWERING LOCUS T decreases the first flower node in bitter gourd (*Momordica charantia* L.). *Front. Plant Sci.* 14:1153208. doi: 10.3389/fpls.2023.1153208

A point mutation in *MC06g1112* encoding FLOWERING LOCUS T decreases the first flower node in bitter gourd (*Momordica charantia* L.)

Jian Zhong[1,2], Junjie Cui[3], Mingjun Miao[2], Fang Hu[4], Jichi Dong[1], Jia Liu[1], Chunfeng Zhong[1], Jiaowen Cheng[1]* and Kailin Hu[1]*

[1]Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (South China), College of Horticulture, South China Agricultural University, Guangzhou, China, [2]Horticulture Research Institute, Sichuan Academy of Agricultural Sciences, Chengdu, Sichuan, China, [3]Department of Horticulture, Foshan University, Foshan, China, [4]Henry Fok School of Biology and Agricultural, Shaoguan University, Shaoguan, China

In Cucurbitaceae crops, the first flower node (FFN) is an important agronomic trait which can impact the onset of maturity, the production of female flowers, and yield. However, the gene responsible for regulating FFN in bitter gourd is unknown. Here, we used a gynoecious line (S156G) with low FFN as the female parent and a monoecious line (K8-201) with high FFN as the male parent to obtain F_1 and F_2 generations. Genetic analysis indicated that the low FFN trait was incompletely dominant over the high FFN trait. A major quantitative trait locus (QTL)-*Mcffn* and four minor effect QTLs-*Mcffn1.1*, *Mcffn1.2*, *Mcffn1.3*, and *Mcffn1.4* were detected by whole-genome re-sequencing-based QTL mapping in the S156G×K8-201 F_2 population (n=234) cultivated in autumn 2019. The *Mcffn* locus was further supported by molecular marker-based QTL mapping in three S156G×K8-201 F_2 populations planted in autumn 2019 (n=234), autumn 2020 (n=192), and spring 2022 (n=205). Then, the *Mcffn* locus was fine-mapped into a 77.98-kb physical region on pseudochromosome MC06 using a large S156G×K8-201 F_2 population (n=2,402). *MC06g1112*, which is a homolog of *FLOWERING LOCUS T* (*FT*), was considered as the most likely *Mcffn* candidate gene according to both expression and sequence variation analyses between parental lines. A point mutation (C277T) in *MC06g1112*, which results in a P93S amino acid mutation between parental lines, may be responsible for decreasing FFN in bitter gourd. Our findings provide a helpful resource for the molecular marker-assisted selective breeding of bitter gourd.

KEYWORDS

bitter gourd, first flower node, quantitative trait locus, fine-mapping, *FLOWERING LOCUS T*

Introduction

The appearance of the first flower is a signal of the pivotal transition from vegetative to reproductive growth in flowering plants (Pnueli et al., 1998; Zahid et al., 2021). Both the time of first flowering and the first flower node (FFN) are useful for the evaluation of crop maturity, and are thus considered important agronomic traits in crop improvement endeavors (Yuan et al., 2008; Zhang et al., 2018; Zhang et al., 2019). In the model plant *Arabidopsis thaliana*, *FLOWERING LOCUS T* (*FT*) is an important regulator gene in determining the flowering time (Corbesier et al., 2007; Turck et al., 2008). The function of the *FT* gene has also been characterized in several Cucurbitaceae crops. In cucumber (*Cucumis sativus*), the *CsFT* gene has been reported to explain 52.3% of the phenotypic variation in flowering time, and is theorized to have been crucial to the spread of this species from its origin in the tropics to higher latitudes (Lu et al., 2014; Wang et al., 2020). The overexpression of *CsFT*, as well as *Cm-FTL1* and *Cm-FTL2* from squash (*Cucurbita maxima*) and *CmFT* from melon (*Cucumis melo*), in *Arabidopsis* promotes early flowering (Lin et al., 2007; Yang et al., 2022). Furthermore, the overexpression of *Arabidopsis*-derived *AtFT* in squash also results in early flowering (Lin et al., 2007).

Several studies have reported that other genes are also associated with the regulation of flowering time in Cucurbitaceae crops. For example, in cucumber, silencing *CsGL2-LIKE* results in delayed male flowering through inhibition of *CsFT* expression (Cai et al. (2020). Yi et al. (2020), utilizing haplotype analysis, report that *ClGA2/KS* is associated with flowering time in watermelon (*Citrullus lanatus*). The overexpression of cucumber-derived *CsTFL1b*, a homolog of *TERMINAL FLOWER 1* (*TFL1*), results in later flowering in transgenic *Arabidopsis* (Zhao et al., 2018; Cai et al., 2020). Contrarily, the overexpression of cucumber-derived *CsBCAT* (*CsBCAT2*, *CsBCAT3*, and *CsBCAT7*) and *CsMADS02* has been shown to accelerate flowering in transgenic *Arabidopsis* (Lee et al., 2019; Zhou et al., 2019). Although the function of these genes has not been universally verified, these initial reports provide clues for the further dissection of the regulation of flowering time in Cucurbitaceae crops. In addition, several quantitative trait loci (QTLs) have been reported to be associated with flowering time in cucurbits. Pan et al. (2017) and Sheng et al. (2020) identified three and two QTLs associated with flowering time in cucumber, respectively. McGregor et al. (2014) identified a major QTL associated with male flowering time in watermelon, which was later verified by Gimode et al. (2020).

Bitter gourd (*Momordica charantia*), so named because of its characteristically bitter taste, is an edible and medicinal cucurbit that has been used to treat hypertension, cancer, diabetes, infection, hyperlipidemia, and obesity (Akihisa et al., 2007; Zhang et al., 2012; Wang et al., 2017). Bitter gourd originated in Africa (Schaefer et al., 2009; Schaefer and Renner, 2010) and has become an important crop across Asia, Africa, the Caribbean, and South America, among other regions (Basch et al., 2003). In bitter gourd, low FFN or early flowering is usually considered as an important indicator of the early maturity trait. To date, genetic mapping studies have revealed at least 21 QTLs associated with female flowering time and 12 QTLs associated with male flowering time in bitter gourd (Wang and Xiang, 2013; Cui et al., 2018; Gangadhara Rao et al., 2018; Kaur et al., 2022). However, there are currently no research reports about genetic mapping of the FFN trait in bitter gourd.

Thanks to the completion of the fully sequenced and assembled bitter gourd genome (Urasaki et al., 2017; Cui et al., 2020; Matsumura et al., 2020), the mapping and cloning of genes controlling important agronomic traits has become easier. Like typical cucurbits species such as cucumber or melon, there are many types of sexual plants in bitter gourd, of which monoecy that carries both unisexual male and female flowers and gynoecy that harbors only female flower have been reported (Kole, 2020; Zhong et al., 2023). Here, we used a segregating F_2 populations crossing from a gynoecious female parent and a monoecious male parent to elucidate the molecular mechanism of FFN regulation in bitter gourd. We first performed a whole-genome re-sequencing-based QTL mapping to rapidly identify FFN-associated genetic loci. Next, we conducted a molecular marker-based classical QTL mapping to confirm the stability of the major effected QTL in three F_2 population cultivated three different environments, respectively. Finally, we fine-mapped the identified candidate gene. Both expression and sequence variation analyses suggest that the candidate gene *MC06g1112* regulates FFN in bitter gourd. The results of this study will be invaluable for breeding improved bigger gourds, and further our understanding of the regulation of floral timing in cucurbits.

Materials and methods

Plant materials

A gynoecious, low-FFN (7-10th nodes) inbred line (S156G, P_1) (Figure 1A) and a monecious, high-FFN (16-19th nodes) inbred line (K8-201, P_2) (Figure 1B) were used as the female and male parents, respectively, to construct the F_1 generation, which was then self-crossed to generate the F_2 population. Both of the parental lines (S156G and K8-201) had been previously whole-genome re-sequenced (Zhong et al., 2022). All plants, representing four generations (P_1, P_2, F_1, and F_2), were cultivated across three quarters (autumn 2019, autumn 2020, and spring 2022) at the experimental field of the SCAU Teaching & Research Base in Zengcheng District, Guangzhou, China (23.24N, 113.64E), under standard agronomic management. Plants from the S156G×K8-201 F_2 population (n=234), which were cultivated in autumn 2019, were used to preliminarily map FFN-associated genetic loci by whole-genome re-sequencing-based QTL mapping. Plants from three S156G×K8-201 F_2 populations, cultivated in autumn 2019 (n=234), autumn 2020 (n=192), and spring 2022 (n=205), were employed for molecular marker-based QTL mapping to confirm the stability of the major effected QTL. Finally, a large S156G×K8-201 F_2 population (n=2,402) was used to fine-map the candidate region associated with FFN. The number of nodes from the node with the first alternate leaf to the node carrying the first flower was used to quantify FFN in bitter gourd.

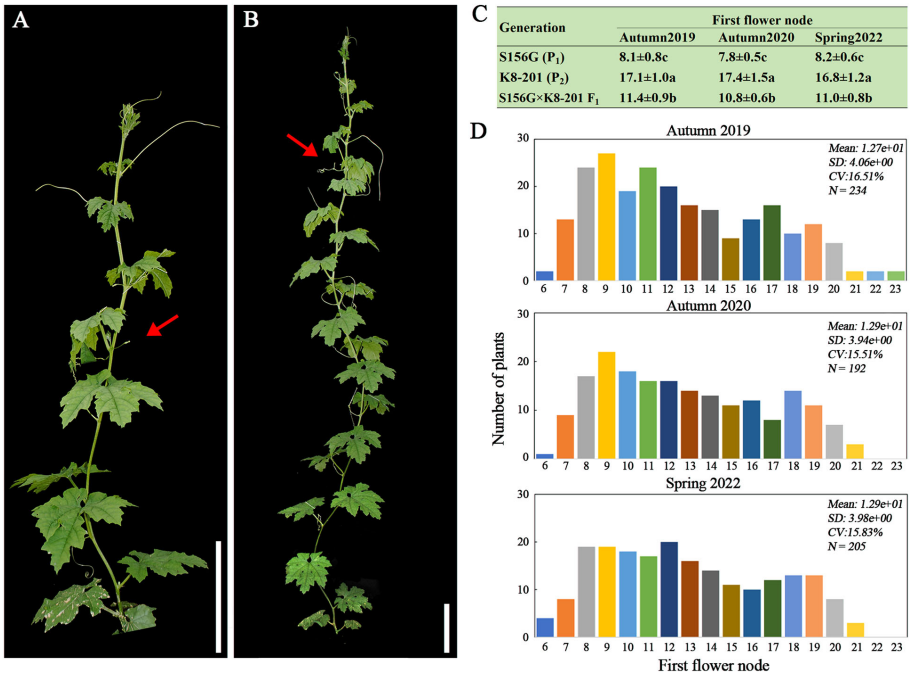


FIGURE 1
Phenotypic evaluation of the FFN trait. **(A)** Low-FFN (7-10th nodes) inbred line S156G (P_1). Bar=10 cm. **(B)** High-FFN (16-19th nodes) inbred line K8-201(P_2). Bar=10 cm. The red arrows in **(A, B)** indicate FFNs. **(C)** Average (± SD) FFN values of the S156G, K8-201, and S156G×K8-201 F_1 generations recorded from autumn 2019, autumn 2020, and spring 2022. Different lowercase letters indicate statistical significance at the 0.01 level. **(D)** Phenotypic distribution of the FFN trait from the three S156G×K8-201 F_2 populations cultivated in autumn 2019, autumn 2020, and spring 2022.

DNA library preparation for whole-genome re-sequencing

The CTAB method (Porebski et al., 1997) was used to isolate genomic DNA (gDNA) from young leaves, and each sample was stored at -20°C prior to analysis. gDNA isolated from the S156G×K8-201 F_2 population (n=234) was utilized to construct DNA sequencing libraries with a KAPA-Hyper Plus Kit (KAPA Biosystems, MA, USA). Briefly, DNA was fragmented by ultrasonication to sizes of 250-350 bp, which were utilized for end-repairing and 3' adenylation. After the adapters were ligated to the ends of these 3'-adenylated fragments, the products were purified by gel recovery. The recovered products were amplified by polymerase chain reaction (PCR) to construct the DNA sequencing libraries. The quality of the DNA sequencing libraries was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and a Real-Time PCR (qPCR) System (Bio-Rad, CA, USA). Finally, the qualified DNA sequencing libraries were sequenced on an Illumina Nova-Seq platform (Illumina, CA, USA).

Whole-genome re-sequencing-based QTL mapping

Quality control of raw whole-genome re-sequencing data, including removal of adapter sequences and low-quality reads, was conducted with Fastp (Chen et al., 2018). Clean reads were aligned to the Dali-11 reference genome with BMA-MEM2 (Vasimuddin et al., 2019; Cui et al., 2020), and the alignment results were evaluated with Qualimap2 (Okonechnikov et al., 2016). Both single nucleotide polymorphisms (SNPs) and insertions and deletions (InDels) were called with BCTtools (Li, 2011), and all variations were annotated with ANNOVAR (Wang et al., 2010). SNPs with a minor allele frequency <0.05 or a missing call frequency >0.1 were removed with VCFtools (Danecek et al., 2011). High quality SNPs were further used to QTL mapping using QTL package in R language. First, the multiple imputation method was used to calculate the LOD value by QTL scanning, and then the significant threshold of LOD value was obtained by 1000 permutation test. Finally, the confidence interval of the selected QTL was identified by LOD support intervals evaluation method.

Molecular marker-based QTL mapping

The variation in SNPs and InDels between the two parental lines (S156G and K8-201) was obtained by aligning the clean re-sequencing data to the Dali-11 reference genome using SOAP2 (Li et al., 2009). Primers for SNP and InDel molecular markers within the whole-genome re-sequencing-based QTL mapping-delimited candidate region were designed with Primer3 Plus (https://www.primer3plus.com), and SNPs were converted to cleaved amplified polymorphic sequences (CAPS) or derived CAPS (dCAPS) markers. Primer sequences are listed in Supplementary Table 1. PCR was carried out in a 10 μL of reaction volume consisting of 0.2 μL of forward and reverse primers (10 μmol/L), 50-100 ng of DNA template, 5 μL of Green Taq Mix (Vazyme, Nanjing, China), and 3.6 μL of nuclease-free water. The PCR

procedure was as follows: initial denaturation at 94 °C for 3 min; 34 cycles of denaturation at 94 °C for 15 s, annealing at 55 °C for 15 s, and extension at 72 °C for 30 s; and final extension at 72 °C for 5 min. The InDel primer-amplified PCR products were directly visualized with 6% polyacrylamide gel electrophoresis (PAGE). The CAPS- or dCAPS-amplified PCR products were first digested with corresponding restriction endonucleases (Supplementary Table 1) at a stationary temperature of 37 °C for 30 min, and the digested products were then visualized with 6% PAGE.

Polymorphic markers from within the molecular marker-based QTL mapping-delimited candidate region were utilized to genotype the three F_2 populations cultivated between autumn 2019, autumn 2020, and spring 2022. Genetic distances of those polymorphic markers were calculated with JoinMap 4.0 (Van Ooijen, 2006). Based on marker genotypes and FFN phenotypes of the three F_2 populations, FFN-associated QTL mapping was conducted with MapQTL 6.0 using the multiple QTL model (MQM mapping) procedure (Van Ooijen, 2009).

Fine-mapping

Recombinant and non-recombinant members of the three F_2 populations were identified using two markers flanking the candidate region identified by molecular marker-based QTL mapping. Non-recombinant plants were divided into three groups (dominant homozygote, recessive homozygote, and heterozygote) depending on whether both flanking markers were identical to S156G, K8-201, or S156G×K8-201 F_1, respectively. During the process of fine-mapping, the average FFN values of the recessive homozygote and heterozygote groups were used as a reference to evaluate the FFN phenotype of the recombinant plant. Recombinant plants were divided into two groups: group one plants contained a recombination of the dominant homozygote and heterozygote genotypes, and group two plants contained a recombination of the recessive homozygote and heterozygote genotypes. Only group two plants were utilized for further genotyping with six newly-developed markers from within the flanked region. By using a combination of FFN phenotype data and genotype markers obtained from the group two plants, we identified a more accurate candidate region and two new flanking markers. These two new flanking markers were used to screen the S156G×K8-201 F_2 population (n=2,402) for plants containing a recombination of the recessive homozygote and heterozygote genotypes. The selected recombinant plants were then grown in the field to evaluate their FFN phenotypes, and genotyped using nine markers from within the newly-identified flanked region. Finally, by using a combination of FFN phenotype data and genotype markers obtained from these recombinant plants, we delimited the FFN-associated fine-mapping interval.

Expression analysis and cloning of the candidate genes

Prior to RNA extraction, tissue samples, including roots, leaves, petioles, female flowers, sarcocarps, and stems (including the 5th, 10th, 15th, 20th, and 25th node [shoot tip, ST]), were collected from parental plants at the 25-leaf stage and frozen in liquid nitrogen. Three biological replicates were used for all analyses. Total RNA was extracted with an Eastep Super Total RNA Extraction kit (Promega, Shanghai, China), and first-strand cDNA was synthesized with an Eastep RT Master Mix kit (Promega, Shanghai, China), according to the manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) was performed using a TB Green Premix Ex TaqTM II kit (Takara Bio, Shiga, Japan) on a CFX384 Real-Time System (Bio-Rad, CA, USA). All primers are listed in Supplementary Table 1. Six categories of tissue samples, including roots, stems (15th node), leaves, petioles, female flowers, and sarcocarps, were utilized to perform qRT-PCR for the genes annotated within the fine-mapped interval. The five categories of stem samples were utilized to perform qRT-PCR for the FNN-associated candidate gene. Three technical replicates were used for all assays. The relative expression level of each gene was normalized using the bitter gourd beta-actin gene (*MC02g1395*) and quantified using the delta-delta Ct method ($2^{-\Delta\Delta Ct}$) (Livak and Schmittgen, 2001).

The primer sequences used to clone the full-length cDNA of the FNN-associated candidate gene were designed according to the gene annotation of Dali-11 reference genome (Cui et al., 2020) (Supplementary Table 1). PCR amplifications of cDNA collected from parental stem (15th node) samples were performed with Phanta Max Super-Fidelity DNA Polymerase (Vazyme, Nanjing, China), according to the manufacturer's instructions. The PCR products were purified and then ligated into the pMD19-T vector (Takara, Shiga, Japan). At least three positive colonies per amplicon were selected for Sanger sequencing, and the generated sequences were assembled with ContigExpress (Lu and Moriyama, 2004). Both nucleotide and amino acid sequences were aligned with ESPript 3.0 (https://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi).

Results

FFN phenotypic characteristics across four generations

We evaluated the FFN phenotype of each plant across four generations, including P_1 (S156G), P_2 (K8-201), F_1, and F_2, planted respectively in autumn 2019, autumn 2020, and spring 2022. Across all three quarters, the FFN of the P_1 (S156G) was significantly lower than that of the P_2 (K8-201) generation, with the FFN of the P_1 (S156G) generation ranging from the 7th to the 9th node (average of ~8th node) and the FFN of the P_2 (K8-201) generation ranging from the 15th to the 19th node (average of ~17th node) (Figures 1A-C). The FFN of the F_1 generation was significantly higher than the P_1 (S156G) generation and lower than the P_2 (K8-201) generation, ranging from the 9th to the 13th node (average of ~11th node) (Figure 1C), indicating that the low FFN trait is incompletely dominant over the high FFN trait. The FFN of plants in the three F_2 populations was highly stratified but tended toward the low FFN of the P_1 (S156G) generation, ranging from the 6th to the 23th node (Figure 1D; Supplementary Table 2).

Whole-genome re-sequencing-based QTL mapping detects FFN-associated genetic locus

Whole-genome re-sequencing of the 234 F_2 individuals planted in autumn 2019 resulted in the generation of 557.9 Gb of raw data, and 533.9 Gb of clean data was obtained after filtering (Supplementary Table 3). The clean data exhibited a Q20 of 97.3% and an average sequencing depth of 7.6×, indicating that the data was of high quality enough for subsequent bioinformatics analysis. Approximately 98.0% of the clean reads were aligned to Dali-11 reference genome, with a sample-specific genomic coverage of 88.6% (Supplementary Table 4). After aligning the clean reads to the Dali-11 reference genome (Cui et al., 2020), a total of 175,019 high-quality SNPs were obtained (Supplementary Figure 1). A QTL mapping combining FFN phenotype and SNP data from the 234 F_2 individuals identified one ~3.77 Mb FFN-associated major effected QTL designated as the *Mcffn* locus located between 9.26 Mb and 13.03 Mb on pseudochromosome MC06 (hereafter referred to as MC06), and four minor effected QTLs named *Mcffn1.1*, *Mcffn1.2*, *Mcffn1.3*, and *Mcffn1.4* located in pseudochromosome MC01, MC02, MC03, and MC08, respectively. (Figure 2A, Supplementary Table 5).

The *Mcffn* locus is narrowed into a 1.61-Mb interval by molecular marker-based QTL mapping

Eleven polymorphic InDel markers (FN1-FN11) were developed within the ~3.77 Mb candidate region (Supplementary Table 1) and used to genotype 631 F_2 individuals cultivated in autumn 2019 (n=234), autumn 2020 (n=192), and spring 2022 (n=205). QTL mapping combining FFN phenotype and marker genotype data revealed that the 11 polymorphic InDel markers exhibited different LOD values between the three quarters: 8.50-53.81 in autumn 2019, 3.27-47.01 in autumn 2020, and 6.90-58.70 in spring 2022 (Figure 2B). These results suggested that all of the 11 InDel markers were linked to the FFN phenotypes. It is worth noting that the maximum LOD values of 56.73, 52.87, and 59.76, which explained 67.3%, 71.9%, and 73.9% of the variation in the FFN phenotype in the three F_2 populations planted in autumn 2019, autumn 2020, and spring 2022, respectively, were all located between two markers, FN5 and FN6 (Figure 2B). Accordingly, we suggested that the *Mcffn* locus was located within a 1.61-Mb physical interval between the FN5 (11,262,463 bp) and FN6 (12,873,591 bp) markers on MC06 (Figure 2C).

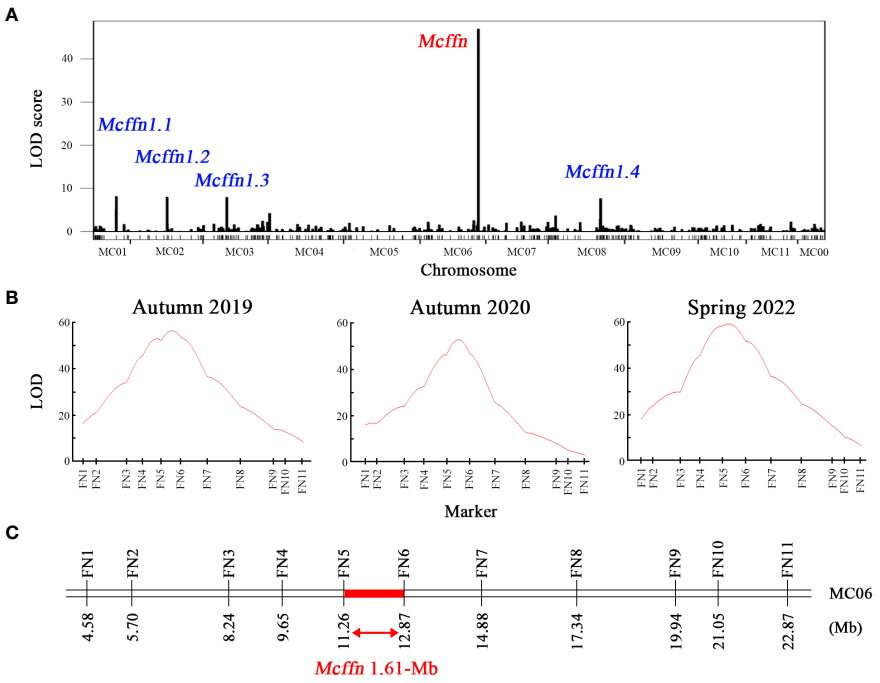


FIGURE 2
Preliminary mapping of the *Mcffn* locus. **(A)** Whole-genome re-sequencing-based QTL mapping in the S156G×K8-201 F_2 mapping population (n=234) planted in autumn 2019. **(B)** Molecular marker-based QTL mapping in the *Mcffn* locus in three S156G×K8-201 F_2 mapping populations planted in autumn 2019, autumn 2020, and spring 2022. **(C)** Physical map of the molecular markers used for molecular marker-based QTL mapping. The numbers under the bar correspond to the physical positions (Mb). The red bar represents the region of the *Mcffn* locus delimited by QTL analysis.

The *Mcffn* locus is fine-mapped into a 77.98-kb interval

Based on genotyping with the two flanking markers (FN5 and FN6), the 631 F_2 individuals were divided into 579 non-recombinant plants and 52 recombinant plants. In order to study the relationship between *Mcffn* genotypes and FFN phenotypes, the 579 non-recombinant plants were divided into three groups: 142 dominant homozygotes, 128 recessive homozygotes, and 309 heterozygotes. The FFN phenotype was significantly different between groups across all three quarters, while within-group differences were not significant (Table 1), suggesting that the FFN trait is genetically, rather than environmentally, determined. To reduce the possibility of errors when categorizing plants as either dominant homozygotes or heterozygotes during the fine-mapping process, we used the FFN value of 11.6 ± 2.6 for the heterozygote genotype and 18.3 ± 2.0 for the recessive homozygote genotype (autumn 2019) as reference criteria (Table 1).

The 52 recombinant plants were divided into two groups: group one plants (n=30) contained a recombination of the dominant homozygote and heterozygote genotypes, and group two plants (n=22) contained a recombination of the recessive homozygote and heterozygote genotypes. The group two plants were further divided into nine haplotypes using six newly-developed markers (FN12-FN17) (Figure 3A). By utilizing the FFN phenotype and marker genotype data of the group two plants, as well as the FFN as the reference, the *Mcffn* locus was further mapped into a 463.01-kb physical interval between the FN13 (11,585,385 bp) and FN16 (12,048,398 bp) markers on MC06 (Figures 3A, B).

To determine a more precise region for the *Mcffn* locus, the S156G×K8-201 F_2 population (n=2,402) was genotyped using the two new flanking markers (FN13 and FN16). Of these, 41 plants containing a recombination of the recessive homozygote and heterozygote genotypes were obtained. Using the FN14 and FN15 markers, and seven newly-developed markers (FN18-FN24), these recombinant plants were divided into 14 haplotypes (Figure 3C and Supplementary Table 1). By utilizing the FFN phenotype and marker genotype data of the 41 recombinant plants, as well as the FFN reference criteria, the *Mcffn* locus was finally fine-mapped into a 77.98-kb physical interval between the FN20 (11,722,144 bp) and FN22 (11,800,118 bp) markers on MC06 (Figures 3C, D).

Differential expression reveals *MC06g1112* as the *Mcffn* candidate gene

By examining the annotation of the Dali-11 reference genome (Cui et al., 2020), we identified four annotated genes (*MC06g1110*, *MC06g1111*, *MC06g1112*, and *MC06g_new0263*) within the 77.98-kb fine-mapping interval. We considered *MC06g1111* a pseudogene because its predicted cDNA is only a 72-bp short nucleotide fragment lacking a complete gene structure and because no transcripts were detected in any sampled tissues (Cui et al., 2020). Only minimal (Cq>35) expression was detected for *MC06g_new0263* across tissues in both parental lines, and no significant differences in the relative expression of *MC06g1110* were detected across tissues between parental lines (Figure 4A). However, *MC06g1112* exhibited significantly different relative expression across tissues between parental lines (Figure 4B). Furthermore, *MC06g1112* exhibited different relative expression across the five different stem categories, increasing from the 5th to the 15th node, and decreasing from the 15th to the 25th node, with almost no expression at the ST (Figure 4C). Additionally, *MC06g1112* exhibited significantly higher expression in the stems of P_1 (S156G) plants than in the stems of P_2 (K8-201) plants at all nodes from the 5th to the 20th (Figure 4C). Accordingly, we proposed that *MC06g1112* was the FFN-associated *Mcffn* candidate gene.

A point mutation of *MC06g1112* may decrease the FFN

By comparing the genomic sequences of the parental lines, we identified seven single-nucleotide variations (SNV-1~SNV-7) within the 77.98-kb fine-mapping interval (Supplementary Table 6). Of these, only SNV-2 (11,775,926 bp) was located within the *MC06g1112* coding region, while the other six SNVs were located in the intergenic spacer region (Supplementary Table 6). The full-length *MC06g1112* cDNA sequences of parental lines were cloned and compared, as a result, *MC06g1112* gene consisted of 540 base pairs, which is a homolog of the *FT* gene encoding a phosphatidylethanolamine-binding protein (PEBP), and therefore was called *McFT*; additionally, we identified a point

TABLE 1 Statistical analysis of FFN trait across three genotypes.

Genotype[a]	Autumn 2019			Autumn 2020			Spring 2020		
	No.	Mean ± SD[e]	SE[g]	No.	Mean ± SD	SE	No.	Mean ± SD	SE
RH[b]	50	18.3 ± 2.0 a[f]	0.28	38	18.4 ± 1.4 a	0.23	40	18.2 ± 1.7 a	0.20
H[c]	112	11.6 ± 2.6 b	0.24	94	12.2 ± 2.3 b	0.24	103	12.2 ± 2.2 b	0.22
DH[d]	54	8.9 ± 1.6 c	0.23	40	8.8 ± 1.5 c	0.24	48	8.5 ± 1.7 c	0.27

[a]Genotype determined by FN5 and FN6,
[b]Recessive homozygote,
[c]Heterozygote,
[d]Dominant homozygote,
[e]The average FFN value ± standard deviation,
[f]Different lowercase letters indicate significance at the 0.01 level,
[g]Standard error.

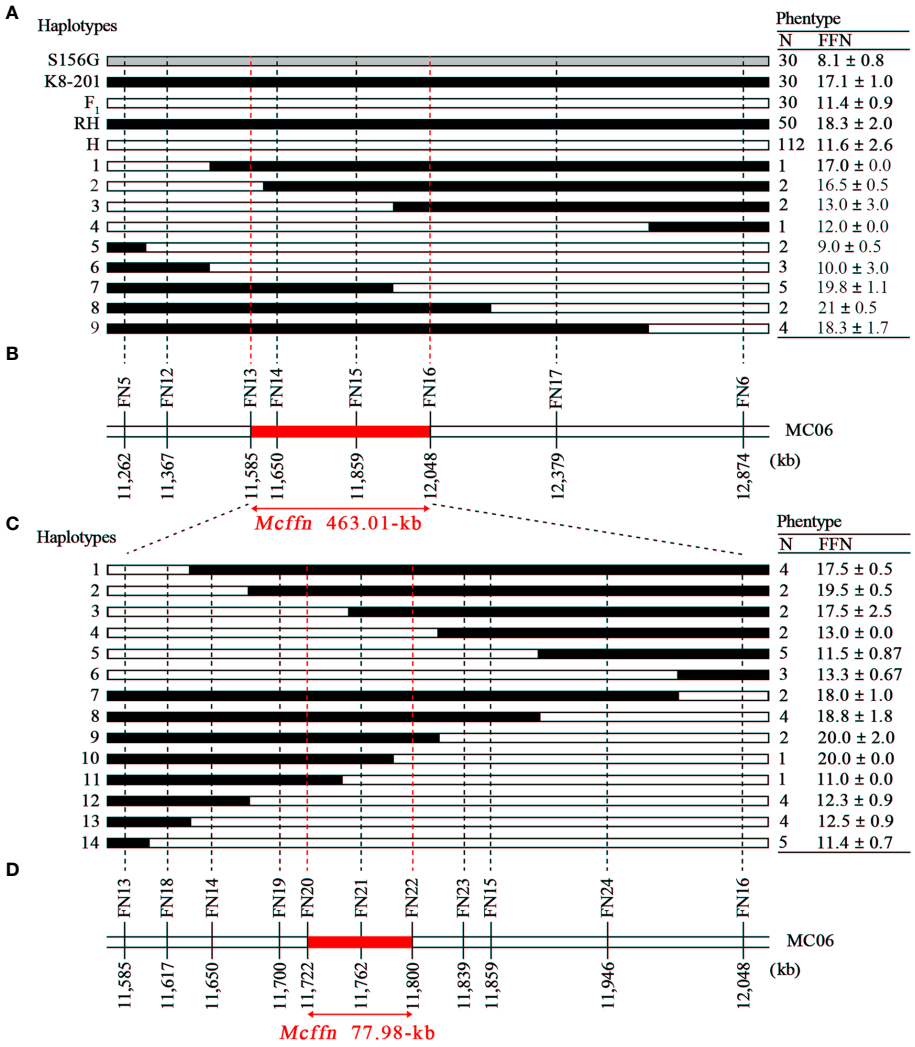


FIGURE 3
Fine-mapping of the *Mcffn* locus. **(A)** Nine haplotypes representing the 22 recombinant plants screened (with flanking markers FN5 and FN6) from the three F_2 populations planted in autumn 2019, autumn 2020, and spring 2022. The dotted red lines indicate the boundaries of the *Mcffn* locus. RH, recessive homozygote. H, heterozygote. N, number. FFN, first flower node. **(B)** Physical map of the molecular markers used to genotype the 22 recombinant plants. The red bar represents the *Mcffn* locus. **(C)** Fourteen haplotypes representing the 41 recombinant plants screened (with flanking markers FN13 and FN16) from the large F_2 population (n=2,402). **(D)** Physical map of the molecular markers used to genotype the 41 recombinant plants (fine-mapping).

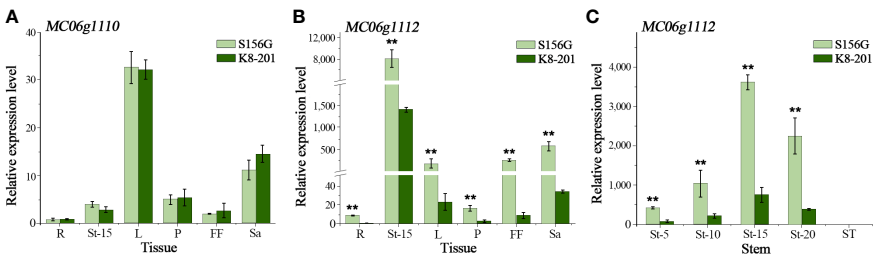


FIGURE 4
Relative expression of candidate genes. **(A)** The relative expression of *MC06g1110* in different tissues at the 25-leaf stage. **(B)** The relative expression of *MC06g1112* in different tissues at the 25-leaf stage. **(C)** The relative expression of *MC06g1112* at different stem node at the 25-leaf stage. R, root. L, leaf. P, petiole. FF, female flower. Sa, sarcocarp. St-5, stem at 5th node. St-10, stem at 10th node. St-15, stem at 15th node. St-20, stem at 20th node. ST, stem at 25th node. The expression levels are presented as the mean ± SD (n=3). ** represents significance at the 0.01 level (Student's t test).

10.3389/fpls.2023.1153208

mutation (C>T) located 277 bp away from the *MC06g1112* start codon, which leaded proline (P) of K8-201 to serine (S) of S156G (hereinafter referred to as P93S) (Figure 5).

To further verify the association of SNV-2 (C277T) and FFN, we designed a dCPAS marker to target SNV-2 by introducing a mismatched base (C) at the end of forward primer to create a *Msp* I restriction enzyme site, which can theoretically produce a 184-bp single fragment with the DNA template of K8-201, a 205-bp single fragment with the DNA template of S156G, and double fragments of 184-bp and 205-bp with the DNA template of S156G×K8-201 F_1 generation (Supplementary Figure 2). Actually, however, the results of marker genotyping showed that both S156G and S156G×K8-201 F_1 generation displayed double fragments of 184-bp and 205-bp, and K8-201 displayed a single 184-bp fragment; in 234 S156G×K8-201 F_2 individuals (autumn 2019), all dominant homozygous and heterozygous plants exhibited double fragments of 184-bp and 205-bp, and all recessive homozygous plants exhibited a single 184-bp fragment (Supplementary Figure 3). Therefore, we conducted Sanger sequencing targeting SNV-2, which indicated that the SNV-2 locus in both DNA and cDNA of K8-201 were recessive homozygous genotype (C), in cDNA of S156G was dominant homozygous genotype (T), while in DNA of S156G was heterozygous genotype (C/T) (Supplementary Figure 4), which implied that the region where SNV-2 is located might have two or multiple copies on the bitter gourd genome.

The McFT proteins of the two parental lines were compared with previously-characterized FT from *C. sativus* (CsFT), *C. melo* (CmFT), *C. lanatus* (ClFT), *Benincasa hispida* (BhiFT), *Lagenaria siceraria* (LsiFT), *Cucurbita maxima* (Cm-FTL1 and Cm-FTL2), *Cucurbita moschata* (Cmo-FTL1 and Cmo-FTL2), *Nicotiana tabacum* (NtFT), *Oryza sativa* (OsFT/Hd3a), and *Arabidopsis thaliana* (AtFT). Sequence alignment and phylogenetic analysis indicated that the McFT proteins from S156G and K8-201 were highly homologous with these previously-characterized FTs, especially FTs of cucurbits species, sharing between 73.63 and 97.21% sequence identity (Figure 6 and Supplementary Figure 5). Additionally, P93 was a strictly-conserved amino acid across all of the examined species, with only one mutant S93 identified in S156G (Figure 6). We speculated that the P93S mutation might be responsible for the decreased FFN exhibited by S156G, since the FFN of S156G was significantly lower than that of K8-201.

Discussion

The onset of flowering, which signals the transition from vegetative to reproduction growth, is a particularly important agronomic trait in Cucurbitaceae crops, as this trait can influence the onset of maturity, the production of female flowers, and yield (Lu et al., 2014; Zhao et al., 2018; Wen et al., 2019). Previous research on the timing of flowering in bitter gourd has primarily focused on the time of onset of flowering, either female or male, from sowing, and has led to the identification of flowering-associated QTLs through genetic mapping (Wang and Xiang, 2013; Cui et al., 2018; Gangadhara Rao et al., 2018; Kaur et al., 2022). However, the gene responsible for regulating flowering time in bitter gourd remained unidentified.

As one of the model plants for research on sex differentiation, Cucurbitaceae species harbor all three basic types of flower sexes, namely female, male, and hermaphroditic flowers (Dellaporta and Calderon-Urrea, 1993; Schaefer and Renner, 2011). All these three basic types carry both pistil and stamen primordia at early development stage of flower bud, and the formation of female and male flowers are resulted by the arrest of stamen and pistil development, respectively (Bai et al., 2004). Previous studies have revealed that the "arrest" processes are genetically controlled, such as the loss of function of *CmWIP1*, *CsWIP1*, and *ClWIP1* lead to gynoecious lines in melon, cucumber, and watermelon, respectively (Martin et al., 2009; Hu et al., 2017; Zhang et al., 2020). Also, our previous works using S156G×K8-201 F_2 population have confirmed that the gene locus responsible for gynoecy in bitter gourd is located at the end of MC01 (Zhong et al., 2023). In addition, our results of

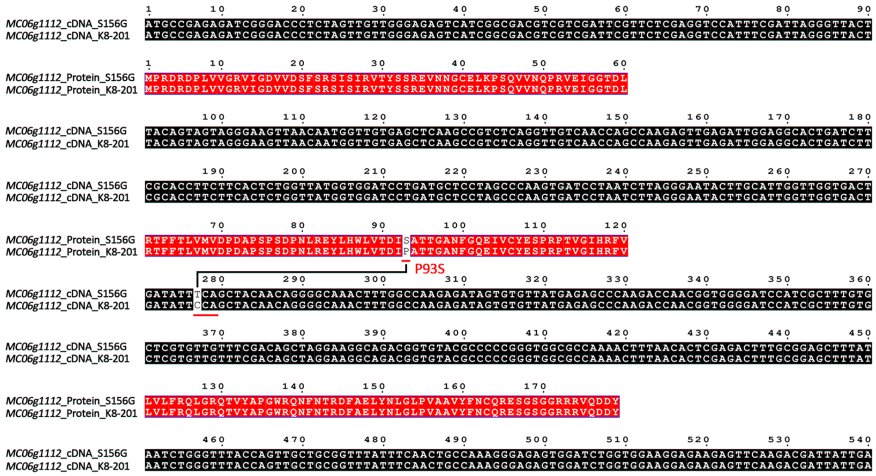


FIGURE 5
Alignment of full-length cDNA and amino acid sequences of *MC06g1112* between S156G and K8-201. Black boxes represent cDNA sequences and red boxes represent amino acid sequences.

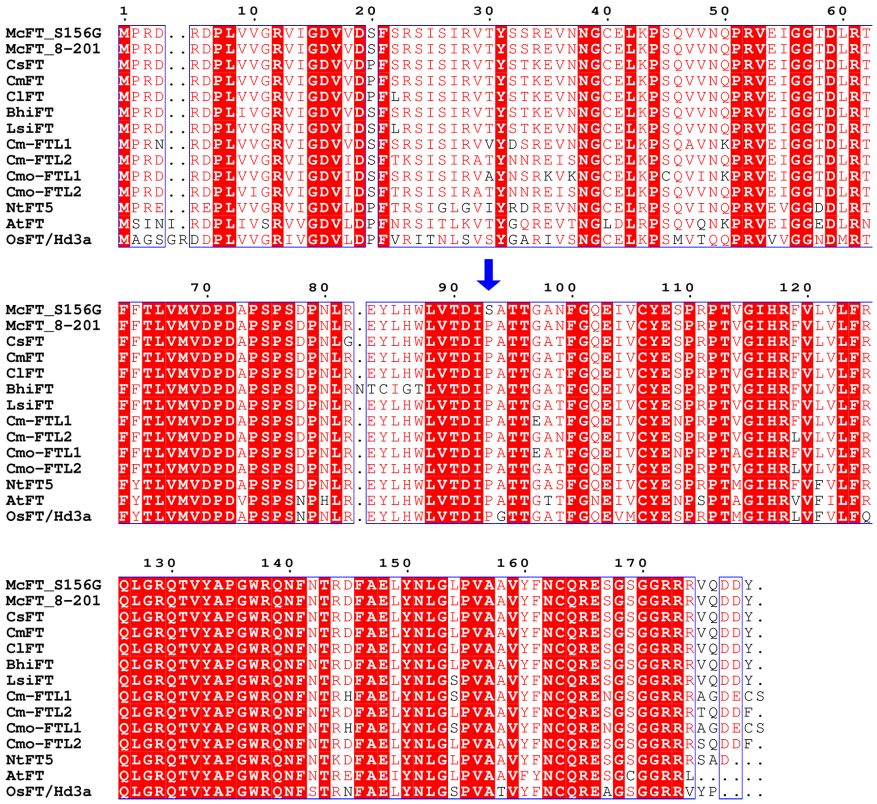


FIGURE 6
Alignment of amino acid sequences of FT proteins across different species of flowering plants. McFT, *Momordica charantia*; CsFT, *Cucumis sativus*; CmFT, *Cucumis melo*; ClFT, *Citrullus lanatus*; BhiFT, *Benincasa hispida*; LsiFT, *Lagenaria siceraria*; NtFT5, *Nicotiana tabacum*; OsFT/Hd3a, *Oryza sativa*; AtFT, *Arabidopsis thaliana;* Cm-FTL1 and Cm-FTL2, *Cucurbita maxima*; Cmo-FTL1 and Cmo-FTL2, *Cucurbita moschata*. The blue arrow indicates the P93S amino acid mutation. The source of each FT protein sequence is listed in Supplementary Table 7.

phenotype investigation showed that there was no direct relationship between gynoecy and FFN in the three S156G×K8-201 F_2 populations (Supplementary Table 2). Hence, we speculate that gynoecy and FFN are independently inherited in bitter gourd.

Here, we used FFN as a proxy for flowering time in bitter gourd and detected a main effect QTL (*Mcffn*) associated with FFN via whole-genome re-sequencing-based QTL mapping (Figure 2A). Then molecular marker-based QTL mapping indicated that *Mcffn* could explain 67.3-73.9% of the variation in the FFN phenotype (Figure 2B), which is higher than the explanatory power reported for QTLs related to either female or male flowering time in bitter gourd (Wang and Xiang, 2013; Cui et al., 2018; Gangadhara Rao et al., 2018; Kaur et al., 2022). Furthermore, the consecutive variation in FFN exhibited by the segregating F_2 populations (Figure 1D) implies that there may be multiple QTLs associated with flowering time in bitter gourd, which is consistent with our QTL mapping results (Figure 2A). Based on fine-mapping, gene expression, and sequence comparison analyses, the *MC06g1112* (*McFT*) gene was identified as the most likely FFN-associated *Mcffn* candidate gene (Figures 3–5). In cucumber, *CsFT* has been found to explain 52.3% of the variation in flower time, and two large structural variations upstream of *CsFT* are associated with earlier flowering (Lu et al., 2014; Zhao et al., 2018; Gimode et al., 2020). Previous comparative genome analyses have shown that most of genomic sequences of pseudochromosome MC06 of bitter gourd, including the genomic fragment where *McFT* is located, are mapped to the chromosome 1 of cucumber (Cui et al., 2020). It is worth mentioning that *CsFT* is just in this collinear genomic interval (Lu et al., 2014), which may imply that bitter gourd and cucumber evolved from the same ancestor and the molecular mechanism regulating FFN or flowering time is highly similar in bitter gourd and cucumber. In squash, the ectopic expression of *Arabidopsis*-derived *AtFT*, which is responsive to inductive short-day (SD) photoperiods, has highly effective in mediating floral induction under long-day (LD) treatment (Lin et al., 2007). These evidences indicate that *FT* genes of cucurbit species may be conserved, and thus its development and application are beneficial to early maturity breeding for cucurbit crops.

Several previous studies have confirmed that the *FT* gene is the downstream target of many transcription factors (TFs) associated with flowering time, such as *CONSTANS* (*CO*), *PHYTOCHROME INTERACTING FACTOR4* (*PIF4*), *FLOWERING LOCUS C* (*FLC*), and *PHYTOCHROME AND FLOWERING TIME 1* (*PFT1*), among others (Putterill et al., 1995; Cerdán and Chory, 2003; Crevillén and Dean, 2011; Kumar et al., 2012), and hence plays a vital role in regulating flowering time across diverse flowering plants, including *Arabidopsis*, rice (*O. sativa*), and winter oilseed rape (*Brassica napus*), among others (Komiya et al., 2008; Ho and Weigel, 2014;

Vollrath et al., 2021). In *Arabidopsis*, the *FT* mRNA is expressed in the vasculature of cotyledons and leaves while the FT protein interacts with a bZIP TF (FD) in the shoot apex to promote floral transition and initiate floral development (Abe et al., 2005; Wigge et al., 2005). Lin et al. (2007), using squash as a model system, report that the FT protein is translocated long distances through the phloem to the shoot apical meristem, where it induces flowering. Because of this, FT is generally considered a long-distance signal, or a leaf-to-apex communicator, for the induction of flowering (Corbesier et al., 2007; Lin et al., 2007). In this study, we detected almost no expression of *McFT* in STs of both parental lines (Figure 4C), suggesting that the functional mechanism of *McFT* in bitter gourd may be similar with previously-reported squash (Lin et al., 2007). In addition, previous studies have focused on *FT* expression in cucumber mainly using leaf tissues (Lu et al., 2014; Wang et al., 2020; Yang et al., 2022). Unlike these previous studies, we examined *FT* expression in several bitter gourd tissues and found that this gene was expressed in all tissues (with the exception of STs), and the expression was particularly high in stem tissues (Figure 4B). Our results suggest that the stem tissue may have the greatest impact on flowering time, although the precise regulatory mechanism underlying *McFT* expression requires further study.

In general, FT is a highly conserved protein which is robust to a wide range of mutations and plays a similar functional role in many species (Lin et al., 2007; Ho and Weigel, 2014; Putterill and Varkonyi-Gasic, 2016). However, Ho and Weigel (2014) reported that the P93 mutation of the FT protein may alter flowering time in *Arabidopsis*, with the P93A and P93T mutations resulting in early flowering and the P93H mutation resulting in delayed flowering. In the present study, we displayed the conservatism of P93 of FT protein across cucurbit species and some other flowering plant species, and identified a P93S amino acid mutation of the McFT protein in bitter gourd (Figure 6). We speculate that the P93S mutation may be responsible for the decreased FFN exhibited by S156G (Table 1). Furthermore, the results of genotypes and Sanger sequencing targeting SNV-2 (C277T) suggested that the region where SNV-2 is located might have two or multiple copies on the bitter gourd genome (Supplementary Figure 3 and 4). The genome replication events may still be the cause of the change of FFN, such as *CsACS1G*, which is a copy of *CsACS1* and leads to gynoecy in cucumber (Mibus and Tatlioglu, 2004; Li et al., 2020). Overall, our results suggest that the *FT* genes may be highly conserved across cucurbits, and thus they should be considered targets for the molecular breeding of early-maturing Cucurbitaceae crops.

Conclusions

FFN of bitter gourd is regulated by a major effect QTL named *Mcffn*, with the low FFN is incompletely dominant over the high FFN. The *Mcffn* locus was fine-mapped into a 77.98-kb physical region on MC06. *MC06g1112*, a homolog of *FT*, was considered as the most likely *Mcffn* candidate gene according to expression and sequence variation analyses. A point mutation (C277T) in *MC06g1112*, which results in a P93S amino acid mutation between parental lines, may be responsible for decreasing FFN in bitter gourd.

Data availability statement

The data presented in the study are deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences repository, accession number CRA009976.

Author contributions

KH and JWC conceived and designed the research. JZ performed most of the experiments and wrote the manuscript. JL, CFZ and MJM performed statistical analysis. JJC, FH and JCD provided helpful discussions. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Science and Technology Planning Project of Guangdong Province (2022B0202160015, 2019A050520002 and 2022-NPY-00-027) and the Guangzhou Science and Technology Plan Project (202206010170 and SL2024A04J01673).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1153208/full#supplementary-material

References

Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., et al. (2005). FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science* 309, 1052–1056. doi: 10.1126/science.1115983

Akihisa, T., Higo, N., Tokuda, H., Ukiya, M., Akazawa, H., Tochigi, Y., et al. (2007). Cucurbitane-type triterpenoids from the fruits of *Momordica charantia* and their cancer chemopreventive effects. *J. Nat. Prod.* 70, 1233–1239. doi: 10.1021/np068075p

Bai, S. L., Peng, Y. B., Cui, J. X., Gu, H. T., Xu, L. Y., Li, Y. Q., et al. (2004). Developmental analyses reveal early arrests of the spore-bearing parts of reproductive organs in unisexual flowers of cucumber (*Cucumis sativus* L.). *Planta* 220, 230–240. doi: 10.1007/s00425-004-1342-2

Basch, E., Gabardi, S., and Ulbricht, C. (2003). Bitter melon (*Momordica charantia*): a review of efficacy and safety. *Am. J. Health Syst. Pharm.* 60, 356–359. doi: 10.1093/ajhp/60.4.356

Cai, Y., Bartholomew, E. S., Dong, M., Zhai, X., Yin, S., Zhang, Y., et al. (2020). The HD-ZIP IV transcription factor GL2-LIKE regulates male flowering time and fertility in cucumber. *J. Exp. Bot.* 71, 5425–5437. doi: 10.1093/jxb/eraa251

Cerdán, P. D., and Chory, J. (2003). Regulation of flowering time by light quality. *Nature* 423, 881–885. doi: 10.1038/nature01636

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Corbesier, L., Vincent, C., Jang, S., Fornara, F., Fan, Q., Searle, I., et al. (2007). FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science* 316, 1030–1033. doi: 10.1126/science.1141752

Crevillén, P., and Dean, C. (2011). Regulation of the floral repressor gene *FLC*: the complexity of transcription in a chromatin context. *Curr. Opin. Plant Biol.* 14, 38–44. doi: 10.1016/j.pbi.2010.08.015

Cui, J., Luo, S., Niu, Y., Huang, R., Wen, Q., Su, J., et al. (2018). A RAD-based genetic map for anchoring scaffold sequences and identifying QTLs in bitter gourd (*Momordica charantia*). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00477

Cui, J., Yang, Y., Luo, S., Wang, L., Huang, R., Wen, Q., et al. (2020). Whole-genome sequencing provides insights into the genetic diversity and domestication of bitter gourd (*Momordica* spp.). *Hortic. Res.* 7, 85. doi: 10.1038/s41438-020-0305-5

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., and DePristo, M. A. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dellaporta, S. L., and Calderon-Urrea, A. (1993). Sex determination in flowering plants. *Plant Cell* 5, 1241–1251. doi: 10.1105/tpc.5.10.1241

Gangadhara Rao, P., Behera, T. K., Gaikwad, A. B., Munshi, A. D., Jat, G. S., and Boopalakrishnan, G. (2018). Mapping and QTL analysis of gynoecy and earliness in bitter gourd (*Momordica charantia* L.) using genotyping-by-sequencing (GBS) technology. *Front. Plant Sci.* 9 1555. doi: 10.3389/fpls.2018.01555

Gimode, W., Clevenger, J., and McGregor, C. (2020). Fine-mapping of a major quantitative trait locus *Qdff3-1* controlling flowering time in watermelon. *Mol. Breed.* 40, 1–12. doi: 10.1007/s11032-019-1087-z

Ho, W. W., and Weigel, D. (2014). Structural features determining flower-promoting activity of *Arabidopsis* FLOWERING LOCUS T. *Plant Cell* 26, 552–564. doi: 10.1105/tpc.113.115220

Hu, B., Li, D., Liu, X., Qi, J., Gao, D., Zhao, S., et al. (2017). Engineering non-transgenic gynoecious cucumber using an improved transformation protocol and optimized CRISPR/cas9 system. *Mol. Plant* 10, 1575–1578. doi: 10.1016/j.molp.2017.09.005

Kaur, G., Pathak, M., Singla, D., Chhabra, G., Chhuneja, P., and Kaur Sarao, N. (2022). Quantitative trait loci mapping for earliness, fruit, and seed related traits using high density genotyping-by-sequencing-based genetic map in bitter gourd (*Momordica charantia* L.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.799932

Kole, C. (2020). *The bitter gourd genome* (Cham Switzerland: Springer).

Komiya, R., Ikegami, A., Tamaki, S., Yokoi, S., and Shimamoto, K. (2008). *Hd3a* and *RFT1* are essential for flowering in rice. *Development* 135, 767–774. doi: 10.1242/dev.008631

Kumar, S. V., Lucyshyn, D., Jaeger, K. E., Alós, E., Alvey, E., Harberd, N. P., et al. (2012). Transcription factor PIF4 controls the thermosensory activation of flowering. *Nature* 484, 242–245. doi: 10.1038/nature10928

Lee, J. H., Kim, Y. C., Jung, Y., Han, J. H., Zhang, C., Yun, C. W., et al. (2019). The overexpression of cucumber (*Cucumis sativus* L.) genes that encode the branched-chain amino acid transferase modulate flowering time in *Arabidopsis thaliana*. *Plant Cell Rep.* 38, 25–35. doi: 10.1007/s00299-018-2346-x

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics* 27, 1157–1158. doi: 10.1093/bioinformatics/btr076

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi: 10.1093/bioinformatics/btp336

Li, Z., Han, Y., Niu, H., Wang, Y., Jiang, B., and Weng, Y. (2020). Gynoecy instability in cucumber (*Cucumis sativus* L.) is due to unequal crossover at the copy number variation-dependent *Femaleness* (*F*) locus. *Hortic. Res.* 7, 32. doi: 10.1038/s41438-020-0251-2

Lin, M. K., Belanger, H., Lee, Y. J., Varkonyi-Gasic, E., Taoka, K., Miura, E., et al. (2007). FLOWERING LOCUS T protein may act as the long-distance florigenic signal in the cucurbits. *Plant Cell* 19, 1488–1506. doi: 10.1105/tpc.107.051920

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Lu, H., Lin, T., Klein, J., Wang, S., Qi, J., Zhou, Q., et al. (2014). QTL-seq identifies an early flowering QTL located near *Flowering Locus T* in cucumber. *Theor. Appl. Genet.* 127, 1491–1499. doi: 10.1007/s00122-014-2313-z

Lu, G., and Moriyama, E. N. (2004). Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform.* 5, 378–388. doi: 10.1093/bib/5.4.378

Martin, A., Troadec, C., Boualem, A., Rajab, M., Fernandez, R., Morin, H., et al. (2009). A transposon-induced epigenetic change leads to sex determination in melon. *Nature* 461, 1135–1138. doi: 10.1038/nature08498

Matsumura, H., Hsiao, M. C., Lin, Y. P., Toyoda, A., Taniai, N., Tarora, K., et al. (2020). Long-read bitter gourd (*Momordica charantia*) genome and the genomic architecture of nonclassic domestication. *Proc. Natl. Acad. Sci. U. S. A.* 117, 14543–14551. doi: 10.1073/pnas.1921016117

McGregor, C. E., Waters, V., Vashisth, T., and Abdel-Haleem, H. (2014). Flowering time in watermelon is associated with a major quantitative trait locus on chromosome 3. *J. Amer. Soc Hortic. Sci.* 139, 48–53. doi: 10.21273/JASHS.139.1.48

Mibus, H., and Tatlioglu, T. (2004). Molecular characterization and isolation of the *F/f* gene for femaleness in cucumber (*Cucumis sativus* L.). *Theor. Appl. Genet.* 109, 1669–1676. doi: 10.1007/s00122-004-1793-7

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Pan, Y., Qu, S., Bo, K., Gao, M., Haider, K. R., and Weng, Y. (2017). QTL mapping of domestication and diversifying selection related traits in round-fruited semi-wild Xishuangbanna cucumber (*Cucumis sativus* L. var. *xishuangbannanesis*). *Theor. Appl. Genet.* 130, 1531–1548. doi: 10.1007/s00122-017-2908-2

Pnueli, L., Carmel-Goren, L., Hareven, D., Gutfinger, T., Alvarez, J., Ganal, M., et al. (1998). The *SELF-PRUNING* gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of *CEN* and *TFL1*. *Development* 125, 1979–1989. doi: 10.1242/dev.125.11.1979

Porebski, S., Bailey, L. G., and Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* 15, 8–15. doi: 10.1007/BF02772108

Putterill, J., Robson, F., Lee, K., Simon, R., and Coupland, G. (1995). The *CONSTANS* gene of Arabidopsis promotes flowering and encodes a protein showing similarities to zinc finger transcription factors. *Cell* 80, 847–857. doi: 10.1016/0092-8674(95)90288-0

Putterill, J., and Varkonyi-Gasic, E. (2016). FT and florigen long-distance flowering control in plants. *Curr. Opin. Plant Biol.* 33, 77–82. doi: 10.1016/j.pbi.2016.06.008

Schaefer, H., Heibl, C., and Renner, S. S. (2009). Gourds afloat: a dated phylogeny reveals an Asian origin of the gourd family (Cucurbitaceae) and numerous oversea dispersal events. *Proc. Biol. Sci.* 276, 843–851. doi: 10.1098/rspb.2008.1447

Schaefer, H., and Renner, S. S. (2010). A three-genome phylogeny of *Momordica* (Cucurbitaceae) suggests seven returns from dioecy to monoecy and recent long-distance dispersal to Asia. *Mol. Phylogenet. Evol.* 54, 553–560. doi: 10.1016/j.ympev.2009.08.006

Schaefer, H., and Renner, S. S. (2011). "Cucurbitaceae. In: kubitzki K (ed) flowering plants eudicots: sapindales, cucurbitales, myrtaceae," in *The families and genera of vascular plants*. (Berlin, Heidelberg: Springer Berlin Heidelberg), 112–174. doi: 10.1007/978-3-642-14397-7_10

Sheng, Y., Pan, Y., Li, Y., Yang, L., and Weng, Y. (2020). Quantitative trait loci for fruit size and flowering time-related traits under domestication and diversifying selection in cucumber (*Cucumis sativus*). *Plant Breed.* 139, 176–191. doi: 10.1111/pbr.12754

Turck, F., Fornara, F., and Coupland, G. (2008). Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Annu. Rev. Plant Biol.* 59, 573–594. doi: 10.1146/annurev.arplant.59.032607.092755

Urasaki, N., Takagi, H., Natsume, S., Uemura, A., Taniai, N., Miyagi, N., et al. (2017). Draft genome sequence of bitter gourd (*Momordica charantia*), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* 24, 51–58. doi: 10.1093/dnares/dsw047

Van Ooijen, J. W. (2006). *Joinmap®4, software for calculation of genetic linkage maps in experimental populations* (Wageningen: Kyazma B.V).

Van Ooijen, J. W. (2009). *MapQTL®6: software for the mapping of quantitative trait loci in experimental populations of diploid species* (Wageningen: Kyazma B.V).

Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). "Efficient architecture-aware acceleration of BWA-MEM for multicore systems," in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. (Rio de Janeiro, Brazil: IEEE), 314–324. doi: 10.1109/IPDPS.2019.00041

Vollrath, P., Chawla, H. S., Schiessl, S. V., Gabur, I., Lee, H., Snowdon, R. J., et al. (2021). A novel deletion in *FLOWERING LOCUS T* modulates flowering time in winter oilseed rape. *Theor. Appl. Genet.* 134, 1217–1231. doi: 10.1007/s00122-021-03768-4

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic. Acids Res.* 38, e164. doi: 10.1093/nar/gkq603

Wang, S., Li, H., Li, Y., Li, Z., Qi, J., Lin, T., et al. (2020). *FLOWERING LOCUS T* improves cucumber adaptation to higher latitudes. *Plant Physiol.* 182, 908–918. doi: 10.1104/pp.19.01215

Wang, S., Li, Z., Yang, G., Ho, C. T., and Li, S. (2017). *Momordica charantia*: a popular health-promoting vegetable with multifunctionality. *Food Funct.* 8, 1749–1762. doi: 10.1039/c6fo01812b

Wang, Z., and Xiang, C. (2013). Genetic mapping of QTLs for horticulture traits in a $F_{2\text{-}3}$ population of bitter gourd (*Momordica charantia* L.). *Euphytica* 193, 235–250. doi: 10.1007/s10681-013-0932-0

Wen, C., Zhao, W., Liu, W., Yang, L., Wang, Y., Liu, X., et al. (2019). CsTFL1 inhibits determinate growth and terminal flower formation through interaction with CsNOT2a in cucumber. *Development* 146, dev180166. doi: 10.1242/dev.180166

Wigge, P. A., Kim, M. C., Jaeger, K. E., Busch, W., Schmid, M., Lohmann, J. U., et al. (2005). Integration of spatial and temporal information during floral induction in *Arabidopsis*. *Science* 309, 1056–1059. doi: 10.1126/science.1114358

Yang, A., Xu, Q., Hong, Z., Wang, X., Zeng, K., Yan, L., et al. (2022). Modified photoperiod response of *CsFT* promotes day neutrality and early flowering in cultivated cucumber. *Theor. Appl. Genet.* 135, 2735–2746. doi: 10.1007/s00122-022-04146-4

Yi, L., Wang, Y., Huang, X., Gong, Y., Wang, S., and Dai, Z. (2020). Genome-wide identification of flowering time genes in cucurbit plants and revealed a gene *ClGA2/KS* associate with adaption and flowering of watermelon. *Mol. Biol. Rep.* 47, 1057–1065. doi: 10.1007/s11033-019-05200-z

Yuan, X. J., Pan, J. S., Cai, R., Guan, Y., Liu, L. Z., Zhang, W. W., et al. (2008). Genetic mapping and QTL analysis of fruit and flower related traits in cucumber (*Cucumis sativus* L.) using recombinant inbred lines. *Euphytica* 164, 473–491. doi: 10.1007/s10681-008-9722-5

Zahid, N., Maqbool, M., Hamid, A., Shehzad, M., Tahir, M. M., Mubeen, K., et al. (2021). Changes in Vegetative and Reproductive Growth and Quality Parameters of Strawberry (*Fragaria* × *ananassa* Duch.) cv. Chandler Grown at Different Substrates. *J. Food Qual.* 2021, 9. doi: 10.1155/2021/9996073

Zhang, J., Guo, S., Ji, G., Zhao, H., Sun, H., Ren, Y., et al. (2020). A unique chromosome translocation disrupting *ClWIP1* leads to gynoecy in watermelon. *Plant J.* 101, 265–277. doi: 10.1111/tpj.14537

Zhang, J., Huang, Y., Kikuchi, T., Tokuda, H., Suzuki, N., Inafuku, K., et al. (2012). Cucurbitane triterpenoids from the leaves of *Momordica charantia*, and their cancer chemopreventive effects and cytotoxicities. *Chem. Biodivers.* 9, 428–440. doi: 10.1002/cbdv.201100142

Zhang, X., Wang, G., Chen, B., Du, H., Zhang, F., Zhang, H., et al. (2018). Candidate genes for first flower node identified in pepper using combined SLAF-seq and BSA. *PloS One* 13, e0194071. doi: 10.1371/journal.pone.0194071

Zhang, X. F., Wang, G. Y., Dong, T. T., Chen, B., Du, H. S., Li, C. B., et al. (2019). High-density genetic map construction and QTL mapping of first flower node in pepper (*Capsicum annuum* L.). *BMC Plant Biol.* 19, 167. doi: 10.1186/s12870-019-1753-7

Zhao, W., Gu, R., Che, G., Cheng, Z., and Zhang, X. (2018). *CsTFL1b* may regulate the flowering time and inflorescence architecture in cucumber (*Cucumis sativus* L.). *Biochem. Biophys. Res. Commun.* 499, 307–313. doi: 10.1016/j.bbrc.2018.03.153

Zhong, J., Cheng, J., Cui, J., Hu, F., Dong, J., Liu, J., et al. (2022). *MC03g0810*, an important candidate gene controlling black seed coat color in bitter gourd (*Momordica* spp.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.875631

Zhong, J., Cui, J., Liu, J., Zhong, C., Hu, F., Dong, J., et al. (2023). Fine-mapping and candidate gene analysis of the *Mcgy1* locus responsible for gynoecy in bitter gourd (*Momordica* spp.). *Theor. Appl. Genet.* 136, 81. doi: 10.1007/s00122-023-04314-0

Zhou, Y., Hu, L., Song, J., Jiang, L., and Liu, S. (2019). Isolation and characterization of a *MADS-box* gene in cucumber (*Cucumis sativus* L.) that affects flowering time and leaf morphology in transgenic *Arabidopsis*. *Biotechnol. Biotec. Eq.* 33, 54–63. doi: 10.1080/13102818.2018.1534556

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 October 2023
DOI 10.3389/fpls.2023.1199210

OPEN ACCESS

EDITED BY
Muhammad Kashif Riaz Khan,
Nuclear Institute for Agriculture and Biology, Pakistan

REVIEWED BY
Guanglong Hu,
Beijing Academy of Agricultural and Forestry Sciences, China
Ali Aslam,
Superior University, Pakistan

*CORRESPONDENCE
Yun Li
gxuliyun@gxu.edu.cn
Pingwu Liu
hnulpw@hainanu.edu.cn

†These authors have contributed equally to this work

RECEIVED 03 April 2023
ACCEPTED 27 September 2023
PUBLISHED 23 October 2023

CITATION
Jiang H, Waseem M, Wang Y, Basharat S, Zhang X, Li Y and Liu P (2023) Development of simple sequence repeat markers for sugarcane from data mining of expressed sequence tags. *Front. Plant Sci.* 14:1199210. doi: 10.3389/fpls.2023.1199210

Development of simple sequence repeat markers for sugarcane from data mining of expressed sequence tags

Huahao Jiang[1†], Muhammad Waseem[2,3†], Yong Wang[1], Sana Basharat[4], Xia Zhang[1], Yun Li[1*] and Pingwu Liu[2,3*]

[1]College of Agriculture, Guangxi University, Nanning, China, [2]School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication), Hainan University, Sanya, China, [3]School of Tropical Agriculture and Forestry (School of Agriculture and Rural Affairs, School of Rural Revitalization), Hainan University, Haikou, Hainan, China, [4]Department of Botany, University of Agriculture Faisalabad, Faisalabad, Pakistan

Sugarcane (*Saccharum* spp. hybrids) is a worldwide acclaimed important agricultural crop used primarily for sugar production and biofuel. Sugarcane's genetic complexity, aneuploidy, and extreme heterozygosity make it a challenging crop in developing improved varieties. The molecular breeding programs promise to develop nutritionally improved varieties for both direct consumption and commercial application. Therefore, to address these challenges, the development of simple sequence repeats (SSRs) has been proven to be a powerful molecular tool in sugarcane. This study involved the collection of 285216 expressed sequence tags (ESTs) from sugarcane, resulting in 23666 unigenes, including 4547 contigs. Our analysis identified 4120 unigenes containing a total of 4960 SSRs, with the most abundant repeat types being monomeric (44.33%), dimeric (13.10%), and trimeric (39.68%). We further chose 173 primers to analyze the banding pattern in 10 sugarcane accessions by PAGE analysis. Additionally, functional annotation analysis showed that 71.07%, 53.6%, and 10.3% unigenes were annotated by Uniport, GO, and KEGG, respectively. GO annotations and KEGG pathways were distributed across three functional categories: molecular (46.46%), cellular (33.94%), and biological pathways (19.6%). The cluster analysis indicated the formation of four distinct clusters among selected sugarcane accessions, with maximum genetic distance observed among the varieties. We believe that these EST-SSR markers will serve as valuable references for future genetic characterization, species identification, and breeding efforts in sugarcane.

KEYWORDS

sugarcane, plant breeding, simple sequence repeats (SSR), SSR loci, unigene annotation

1 Introduction

Sugarcane (*Saccharum* spp. hybrids) is a global economic and energy crop, with China ranking third in sugar production. This perennial herb is known for its photosynthetic efficiency, higher biomass accumulation, aneuploid polyploidy (≥8), and genetic heterogeneity (Cordeiro et al., 2003). Conventional sugarcane breeding, primarily by stem cutting is laborious and time-consuming, often taking decades to produce new varieties. The complex genetic background of sugarcane cultivars was derived from interspecific hybridization of *S. spontaneum* L. and *S. officinarum* L. (Garsmeur et al., 2018). The commercial sugarcane cultivars inherit ~70–80% of their chromosomes from *S. officinarum*, ~10–15% from *S. spontaneum*, and the remaining ~5–10% from interspecific recombination (D'hont et al., 1996). The limited introgression in sugarcane breeding has led to a narrow genetic basis in commercial cultivars (Singh et al., 2013).

Simple Sequence Repeats (SSRs) are highly polymorphic short tandem repeats (1 – 6 bp) of nucleotide sequences ubiquitous in the genomes of both eukaryotic and prokaryotic organisms (Tóth et al., 2000). SSRs offer several advantages including transferability between species, co-dominance, minimal expertise, instrumentation dependencies, and reproducibility (Mccouch et al., 2001; Cai et al., 2019). They are widely used in genetic diversity studies (Biswas et al., 2020), population structure analysis (Zalapa et al., 2012), association mapping (Gyawali et al., 2016), and linkage mapping (Sugita et al., 2013). The International Sugarcane Microsatellite Consortium (ISMC) has curated 221 SSR markers in sugarcane cultivar R570 (French Reunion) and Q124 (Australian) (Oliveira et al., 2009). Wu et al. (2019) developed an additional 226 SSR markers using a combined fluorescence-labeled SSR and a high-performance capillary electrophoresis (HPCE) system for parental germplasm of the sugarcane breeding programs in China. Similarly, You et al. (You et al., 2015) successfully employed expressed sequence tag-SSR (EST-SSR) to establish the relationship among 69 varieties of *Colocasia esculenta*. Chen et al. (Chen et al., 2017) characterized 11 varieties of *Lycium* by EST-SSRs. Ukoskit et al. (2019) identified 185 EST-SSRs in cultivated sugarcane "Phil6607" and *S. spontaneum* "S6". Recently, Xiao et al. (Xiao et al., 2020) identified 46,043 SSRs in the diverse panel of sugarcane (22 accessions).

Genome sequencing revolutionized the discovery and application of SSRs in various plant species, including sugarcane. The release of the *S. spontaneum* genome in 2018 (Zhang et al., 2018) has provided a valuable resource for sugarcane cultivar breeders. Previous efforts have yielded a relatively small number of SSR markers in sugarcane, for instance, 351 EST-SSRs were identified from 4085 EST sequences (Singh et al., 2013), 406 EST-SSR markers with 63 were verified as polymorphic (Ul Haq et al., 2016), and 2005 markers were identified from EST sequences with 65.5% showed polymorphism (Oliveira et al., 2009). Therefore, the development of markers to assess the genetic relationships with a comprehensive set of EST information has become an imperative task. In this study, we attempt to screen EST-SSR based on sugarcane unigenes, particularly those associated with functional genes, and assess the genetic diversity among other sugarcane accessions (10 in total) that have been previously overlooked. Additionally, we also investigate the evolutionary relationship between the sugarcane genome to those of sorghum and maize. We believe that these newly developed EST-SSR markers will provide a valuable reference for sugarcane breeding programs and facilitate species screening and identification.

2 Materials and methods

2.1 Plant materials

A panel of diverse sugarcane accessions including wild type and eight cultivated were sources from Guangxi, Yunnan, Taiwan, and Fujian. These accessions were maintained at Guangxi University, Nanning, China (Table 1).

TABLE 1 Information of sugarcane accessions.

Accession Number	Type	Genotypes	Origin	Pedigree
1	Wild	*S. officinarum*	Guangxi	Unknown
2	Wild	*S. spontaneum*	Guangxi	Unknown
3	Cultivated	Yunrui05-782	Yunnan	Hybrid of wild species
4	Cultivated	Yunrui05-767	Yunnan	Hybrid of wild species
5	Cultivated	ROC10	Taiwan	ROC5 × Taitang152
6	Cultivated	ROC22	Taiwan	ROC5 × 69-463
7	Cultivated	Guitang28	Guangxi	CP80-1018 × CP88-2032
8	Cultivated	Guitang32	Guangxi	Yuenong73-204 × CP67-412
9	Cultivated	Funong40	Fujian	Funong93-3406×Yuetang91-976
10	Cultivated	Funong39	Fujian	Yuetang91-976 ×CP84-1198

2.2 EST retrieval and mining

The raw EST sequences (approximately 285216) of sugarcane were downloaded from the NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/dbEST/, on January 14, 2013). The raw sequences were cleaned to remove the poly A (5′ or 3′ end) or poly T stretches using EST-Trimmer software (http://pgrc.ipk-gatersleben.de/misa/download/est_trimmer.pl). Subsequently, we assembled the EST sequences using Contig Assembly Program 3 (CAP3, http://doua.prabi.fr/software/cap3) DNA sequences assembly program, with parameter set as 90% identity and 40 bp overlap.

2.3 Identification of SSR motifs and primer pair design

The assembled EST sequences were subjected to a search for SSR motif using the Microsatellite program (MISA; http://pgrc.ipk-gatersleben.de/misa/) with default parameters as follows: 10 for monomeric repeats, 6 for dimeric repeats, and 5 for trimeric, tetrameric, pentanucleotide, and hexameric repeats each. Subsequently, the primer pair was designed in the program Primer 3.0 with the standard criteria as a primer size of 18 to 27 bp and approximately 20 bp, PCR product size of 100 to 300 bp, GC content from 40 – 60%, and melting temperature (Tm) variation from 57 - 63°C.

For each SSR locus, we selected three primer pairs, and the pair yielding the highest-scoring DNA was selected for subsequent SSR marker studies. *In-silico* PCR analysis of the SSR primer pair was performed using MFEprimer3.2.6 (https://mfeprimer3.igenetech.com/) with default parameter setting, except the Tm was set to 50°C (Yu and Zhang, 2011). The primers were synthesized from Sangon Biotech (Shenzhen, China).

2.4 Genomics DNA extraction and SSR genotyping

The genomic DNA (gDNA) was extracted from young sugarcane leaves using the cetyltrimethylammonium bromide (CTAB) method. A Nanodrop spectrophotometer (thermos Scientific, USA) was used for gDNA quantification followed by 1% agarose gel electrophoresis for the quality of gDNA. Finally, the DNA was normalized to 10 ng μL^{-1} for PCR amplification. The PCR reaction was performed in a total reaction volume of 10 uL containing 30–50 ng of gDNA, 2.0 µL of 10×Taq buffer (Mg2+), 0.2 mM each of dNTPs, 0.5 µM each forward and reverse primer, and 0.5 U of Taq DNA polymerase (Clontech, Takara, Shanghai). The resulting PCR products, along with a 2000 bp DNA marker, were separated on an 8% polyacrylamide gel through electrophoresis and visualized using silver staining.

SSR genotyping data were recorded as one (band present) and zero (band absent). The Polymorphism Information Content (PIC) values were computed using the following formula:

$$PIC = 1 - \sum_{i=1}^{n} Pg2$$

where Pg represents the frequency of a unique genotype if each SSR marker represents a single locus with n SSR genotypes.

The presence and absence of SSR genotyping data were used to construct the phylogenetic tree of 10 sugarcane accessions using the Neighbor-joining (NJ) method based on Nei's genetic distance with the MEGAX program.

2.5 Unigenes annotation in sugarcane and comparison with sorghum and maize

We annotated all the unigenes containing SSRs against Gene Ontology (GO, http://www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) databases. To assess the conservation of sugarcane unigenes, we conducted BLASTN searches against the sorghum (Z3116) and maize (B73) genomes, using an e-value threshold of -15 for sorghum and -10 for maize. Our selection criteria included a sequence identity of more than 80% and a sequence length exceeding 100bp.

3 Results

3.1 Distribution of SSR markers

For SSR analysis, a dataset of 285,216 EST sequences retrieved from the NCBI was subjected to quality and redundancy by the CAP3 program. A total of 23666 unigenes sequences including 4547 contigs were generated (Supplementary Table S1). The unigenes' length ranged from a minimum of 101 bp to a maximum length of 4040 bp, with approximately 17467 unigenes' length varying between 600 to 1200 bases, and 826 unigenes measuring 1800-2400 nucleotides in length (Figure 1A). A summary of the sequencing results is presented in Table 2. Using the MISA identification tool, we predicted 4120 unigenes containing 4960 SSRs, with a frequency of one SSR/4.43 kb of the available ESTs. Among these sequences, 685 ESTs contained more than one SSR, with 415 being compound SSRs featuring multiple types of repeat motifs.

SSR motifs in the *S. spontaneum* genome were found to be highly frequent within gene regions (Figure 2 and Table S1). Of the 4960 SSR loci, we predicted a total of 133 motif types. Analyzing the abundance of SSR types in sugarcane ESTs, we found that monomeric (44.33%), dimeric (13.10%), and trimeric (39.68%) were the most abundant, followed by tetrameric (1.62%), pentameric (0.58%), and hexameric (0.65%) repeat types (Figure 1B and Table 3). Furthermore, we observed that the majority of the SSRs had a length of less than 20 bp, with SSRs between 5 – 7 bp and 10 – 12 bp accounting for 75.90% of all the SSRs identified. Additionally, the number of motif types for monomeric, dimeric, trimeric, tetrameric, pentameric, and hexameric were 2, 6, 30, 47, 22, and 26, respectively

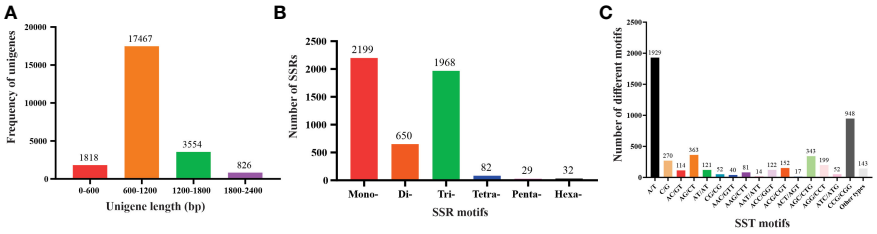


FIGURE 1
Characteristics of unigenes and SSR from the *Saccharum spontaneum* ESTs **(A)** The number of unigenes based on the number of nucleotides in each, **(B)** the number of SSR motifs in monomeric, dimeric, trimeric, tetrameric, pentameric, and hexameric, and **(C)** the number of different SSR motifs in the unigenes of *Saccharum spontaneum*.

(Supplementary Table S2). The frequency distribution of identified SSRs is shown in Figure 1C and Supplementary Table S3. Notably, the most copious motif was A/T (1929) followed by CCG/CGG (948), AG/CT (363), AGC/CTG (343), C/G (270), AGG/CCT (199), ACG/CGT (152), ACC/GGT (122), AT/AT (121), AC/GT (114), AAG/CTT (81), ATC/ATG (52), CG/CG (52), AAC/GTT (40), ACT/AGT (17), and AAT/ATT (14). The remaining 67 motif types had a total count of 143 (Figure 1C and Supplementary Table S3).

Among the monomeric repeats, the A/T motif was the most abundant accounting for 88% of all mono repeats (Figure 3A). For dimeric repeats, the AG/CT motif dominated, constituting 56% of dimeric repeats, followed by AT/AT (19%), AC/GT (17%), and CG/CG (8%) motif types (Figure 3B). In trimeric repeats, CCG/CGG was the most frequent repeat motif, accounting for 48% of trimeric repeats, followed by CGC/CTG (17%), AGG/CCT (10%), ACG/CGT, and ACC/GGT (each at 8%) (Figure 3C). Within tetrameric repeats, AGGC/CCTG, AGGG/CCCT, and ATCC/ATGG (10%) were the most abundant repeat motifs, followed by AAAG/CTTT (8%), and AGAT/ATCT (6%). However, 41% of other types of repeats were also detected in tetrameric repeats (Figure 3D). Within Pentameric repeats, AAAAG/CTTTT (21%) was the most abundant repeat motif, followed by ACAGG/CCTGT (14%), AAGGG/CCCTT (10%), and AGAGG/CCTCT (10%) (Figure 3E). Regarding hexameric repeat, AACATG/ATGTTC (7%) was the most plentiful motif. Other hexameric repeats included AAGCCG/CGGCTT, ACCAGC/CTGGTG, AGAGGG/CCCTCT, and AGGCGG/CCGCCT each accounting for 6%. Additionally, approximately 69% of hexameric repeats were grouped as other types of repeats (Figure 3F).

TABLE 2 Details of ESTs and SSRs identified in sugarcane.

Parameters	Numbers
Total raw EST-sequences	285,216
Contig	4547
Total number of sequences examined	23666
Total size of examined sequences (bp)	22487037
Minimum length of unigenes (bp)	101
Maximum length of unigenes (bp)	4040
Total number of identified SSRs	4960
Number of SSR containing sequences	4120
Number of sequences containing more than 1 SSR	685
Number of SSRs present in compound formation	415
Number of primers designed	3632
Monomeric repeats	2199
Dimeric repeats	650
Trimeric repeats	1968
Tetrameric repeats	82
Pentameric repeats	29
Hexameric repeats	32
Number of motif types	133

3.2 Conservation of SSR in maize and sorghum

To study the evolutionary relationship among sugarcane, maize, and sorghum species and identify unique motifs, we analyzed each motif for the presence of other species. The results showed that sugarcane unigenes were aligned with 11049 unigenes (68.97% of sugarcane unigenes) in sorghum. Of these unigenes, 9382 unigenes were anchored at a single locus, 1002 at two loci, 256 at three loci and four loci, and more for the remaining 409 unigenes. This distribution corresponds to ratios of 84.91%, 9.07%, 2.32%, and 3.70%, respectively. Similarly, 8516 alignments (53.16% of sugarcane unigenes) were revealed between sugarcane and maize; among these unigenes, 4806 mapped to a single locus, 2479 to two loci, 477 to three loci, and 754 unigenes to four loci or more. This distribution corresponds to ratios of 56.43%, 29.11%, 5.60%, and 8.85%, respectively. These results indicate a closer evolutionary relationship between sugarcane and sorghum than that between sugarcane and maize.

3.3 Validation and polymorphisms of SSR primers

The results from *in-silico* PCR analysis showed that 235 of 240 SSR primer pairs had potential amplicons in at least one of the three

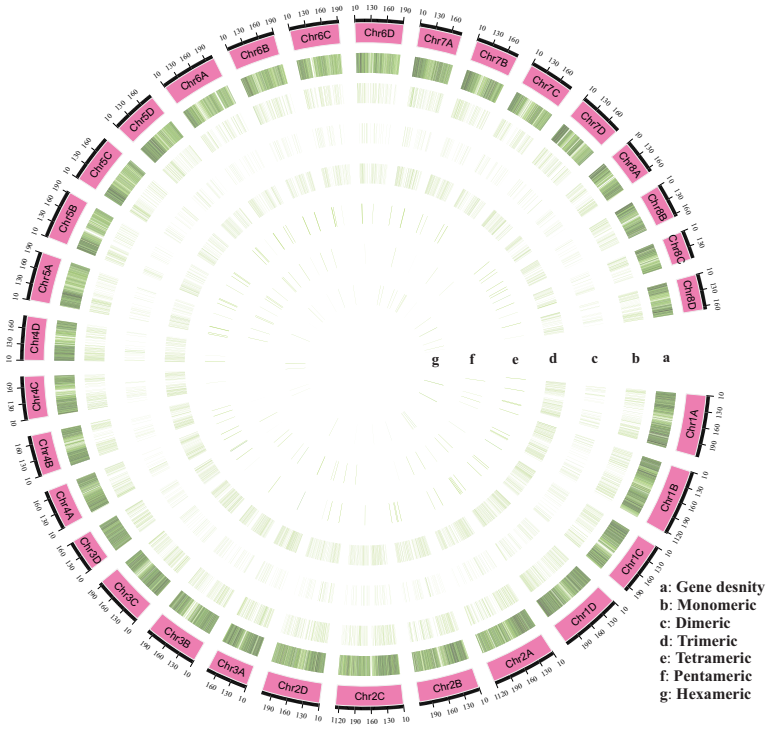


FIGURE 2
An overall view of the distribution of SSR motifs in the chromosomes (Chr01A-Chr08D) of *Saccharum spontaneum* reference genome. **(A)** Gene density, **(B)** Monomeric, **(C)** Dimeric, **(D)** Trimeric, **(E)** Tetrameric, **(F)** Pentameric, and **(G)** Hexameric.

sequenced species including *S. spontaneum*, maize, and sorghum. Interestingly, five of the 242 SSR primer pairs failed to produce potential amplicon in any of these species. Besides, we observed that 9, 13, and 7 SSR primer pairs exclusively generated potential amplicons in the genomes of *S. spontaneum*, maize, and sorghum genome, respectively. Furthermore, 46 SSR primer pairs had potential amplicons in both maize and sorghum genomes, while 9 SSR primer pairs shared potential amplicons in both *S. spontaneum* and sorghum genomes. It is noteworthy that 18 SSR primer pairs were found to have potential amplicons both in *S. spontaneum* and maize genomes. Astonishingly, 133 SSR primer pairs were observed to have potential amplicons in all three genomes (Supplementary Table S4A).

Subsequent analysis of the predicted SSR motifs within the potential amplicons generated by SSR primer pairs showed that 219 of 235 SSR primer pairs had the predicted SSR motif in potential amplicons. Of these, 16 SSR primer pairs only existed in both *S. spontaneum* and maize, while 19 SSR primer pairs were found in the sorghum genome. Similarly, 34 SSR primer pairs had SSR motifs present in both maize and sorghum genomes, and 17 SSR primer pairs shared SSR motifs in both *S. spontaneum* and sorghum genomes. Additionally, 21 SSR primer pairs showed SSR motifs in both *S. spontaneum* and maize genomes. In contrast, 106 SSR primer pairs presented SSR motifs in all three genomes (Supplementary Table S4A).

Among 235 primer pairs with potential amplicons, 40 SSR primer pairs showed at least one base of the primer sequence that did not match with the amplicon. Further analysis of the binding sites of SSR primer pairs with the potential amplicons showed that 53, 10, and 17 SSR primer pairs fully match with at least one of the potential amplicons in *the S. spontaneum*, maize, and sorghum genome, respectively. Nine SSR primer pairs were found to fully match in both maize and sorghum genomes. Thirty-two SSR primer pairs showed full matches in both *S. spontaneum* and sorghum genomes, and 20 SSR primer pairs fully matched in both *S. spontaneum* and maize genomes. Intriguingly, 54 SSR primer pairs were found to fully match in all three genomes (Supplementary Table S4A).

For the applicability of the deduced SSR markers, we selected 173 primer pairs for the analysis in 10 sugarcane accessions including maize and sorghum using PAGE analysis (Figure 4). After optimization, we retained 163 of 173 primers due to clear banding patterns and ease of identification. Among these, 4 were monomeric, 16 were dimeric, 125 were trimeric, 4 were tetrameric, single pentameric, and 3 were hexameric with length ranges spanning from 21 to 109 bp. These 163 SSR loci were capable of amplifying 3-21 alleles within selected accessions, with an average of 9.46 alleles per locus. These SSR markers can be used effectively in genetic diversity analysis, population genetics, and germplasm identification. The polymorphism information content (PIC) values for these SSR loci range from 0.292 to 0.972, with an average PIC value of 0.808, indicating a high level of genetic diversity (Supplementary Table S4B).

Additionally, we gained more insights by integrating *in-silico* PCR analysis and amplification of three primer pairs for each locus. We detected expected PCR products containing SSR loci in both *in-*

10.3389/fpls.2023.1199210

TABLE 3 Summary of frequencies of different SSR repeat motif types.

SSR motif	Number of motif types	5	6	7	8	9	10	11	12	13	14	15	>15	total	Proportion (%)
Mono-	2	0	0	0	0	0	729	348	238	148	110	119	507	2199	44.33
Di-	6	0	258	118	79	38	26	17	14	14	10	6	70	650	13.10
Tri-	30	1264	421	177	60	20	11	5	2	3	3	0	2	1968	39.68
Tetra-	47	58	13	3	2	0	0	1	2	1	1	0	1	82	1.65
Penta-	22	21	4	1	1	0	0	1	1	0	0	0	0	29	0.58
Hexa-	26	19	6	5	0	0	1	1	0	0	0	0	0	32	0.65
Total	133	1362	702	304	142	58	767	373	257	166	124	125	580	4960	

silico PCR analysis and PCR amplification for all three primer pairs in sugarcane. Notably, unexpected PCR bands were amplified for three primer pairs. However, potential amplicons with long fragments, especially more than 1000 bp, were not amplified in maize and sorghum (Table 4 and Figure 4). Additionally, the 265/266 bp bands amplified with primer PW2-23 fully matched in sugarcane and maize were amplified successfully, while the partially matched potential amplicon of 265 bp in sorghum was also amplified. However, for primer pairs PW2-28 and PW2-29, no potential amplicon of the expected PCR products was found in maize and sorghum (Table 4). Nonetheless, an almost identical PCR pattern to sugarcane was observed in maize and sorghum (Figure 4).

3.4 Functional annotation of sugarcane unigenes harboring the SSRs

To explore the potential functions of SSR-containing unigenes, all of these unigenes were annotated against the publicly available functional databases. This analysis indicated that 38.75% of unigenes were associated with GO, while 43.96% were linked to the KEGG. These SSR-containing unigenes were further classified into three major GO functional categories including, biological process, cellular component, and molecular function (Figure 5A and Supplementary Table S5A). Within biological processes, unigenes related to post-embryonic development, photosynthesis, fruit ripening, DNA metabolic process, flower development, and regulation of molecular function accounted for the largest proportion. The cellular component category primarily represented unigenes involved in peroxisome, cytoskeleton, and mitochondrion. In the molecular function category, the most enriched unigenes were involved in signaling receptor activity, protein binding, structural molecule activity, and transporter activity binding.

Furthermore, these unigenes annotated 195 KEGG metabolism pathways, which were classified into six categories including cellular processes, environmental information processing, genetic information processing, metabolism, organismal systems, and brite hierarchies (Figure 5B). In the second level of the pathway classification, prominent categories included carbohydrate metabolism, translation, signal transduction, transport and catabolism, environmental adaptation, protein families: genetic information processing, and protein families associated with signaling and cellular processes. Additional details of each category are provided in Figure 5B and Supplementary Table S5B.

3.5 Genetic diversity and relationships among genotypes

To explore the genetic similarity of sugarcane accession, we conducted a cluster analysis based on a matrix for the presence and absence of deduced alleles. Figure 5 represents the clustering results in the form of phylogenetic trees. The phylogenetic clustering unveiled four distinct accession clusters: "*S. robustum*",

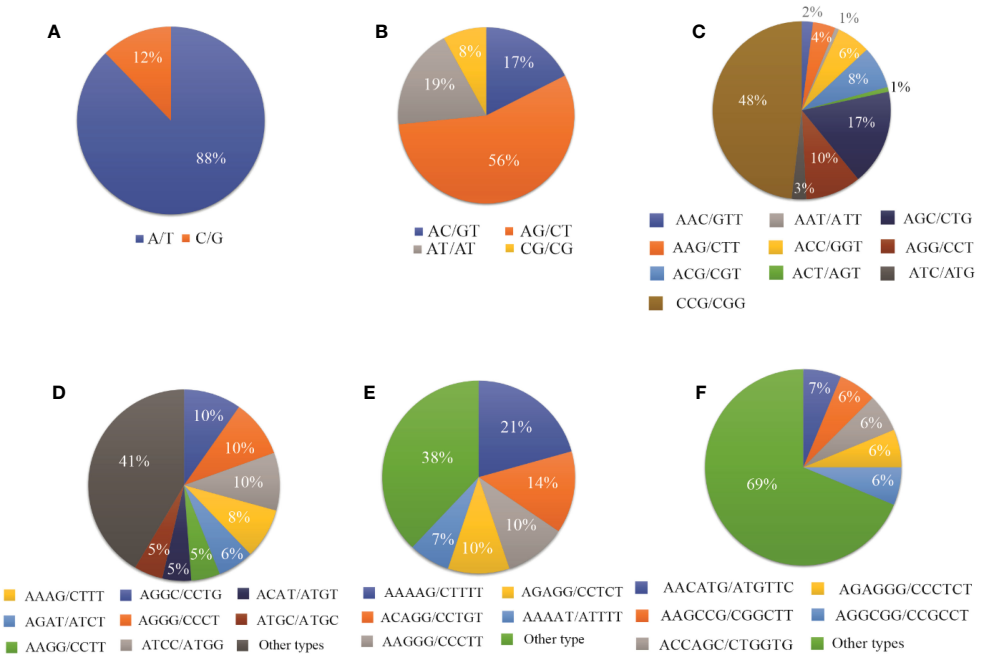


FIGURE 3
The proportion of different repeat motif types in **(A)** Monomeric, **(B)** Dimeric, **(C)** Trimeric, **(D)** Tetrameric, **(E)** Pentameric, and **(F)** Hexameric.

"Yunrui05-767", "ROC10", "ROC22", "Guatang28", and "Guatang32" form a major cluster; Cluster-I, "Funong40", and "Funong39" are present in Cluster-II. "Yunrui05-782" and "*S. spontaneum*" formed a separate cluster each (Cluster-III and IV) at the bottom of the phylogenetic tree (Figure 5). The accession in Cluster-I shares a genetic distance value of 7.4 in relation to other accessions in Cluster-II. Notably, the Taiwan accessions, "ROC22" and "ROC10", as well as the Fujian varieties, "Funong40" and "Funong39" showed a genetic distance of 2.5 between them, indicating a higher degree of similarity as determined by the studied SSR markers. The largest genetic distances were recorded between Yunnan varieties clustered in different clades.

4 Discussion

SSRs are known for their repeatability and polymorphism, extensively being used in unveiling the genetic diversity and markers of assisted breeding programs (Wang et al., 2010) of various plant species including cucumber, cotton, foxtail millet, rice, citrus, horse gram, maize, and sweet cane (Zhou et al., 2021). However, the application of EST-SSR markers has been limited in Sugarcane (*Saccharum* spp.) (Zhang et al., 2012). For instance, Xiao et al. (2020) identified a set of 349 EST-SSR markers. In this study, we have significantly expanded these marker resources by developing a novel set of 4960 EST-SSR markers. Among these

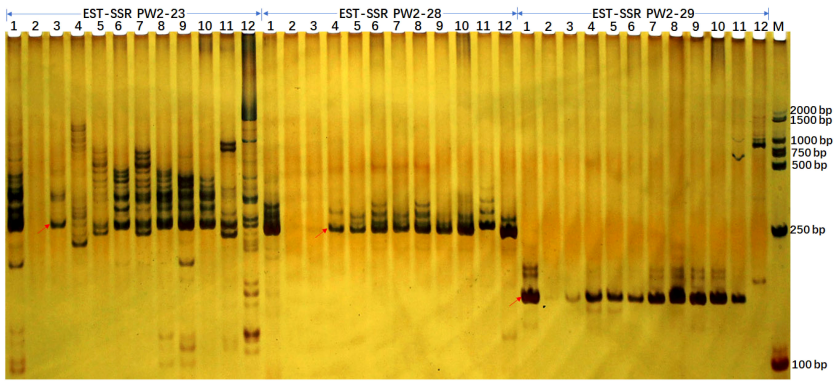


FIGURE 4
EST-SSR verification profiles of 10 accessions from *Sacchraum*, and single accessions from maize and sorghum each detected by polyacrylamide gel electrophoresis. The EST-SSR profiles with PW2-23, PW2-28, and PW2-29 primer pairs were visualized by silver staining. Lanes 1 to 12 were Yunrui05-782, Yunrui05-767, *S. robustum*, *S. spontaneum*, ROC22, ROC10, Guitang28, Guitang32, Funong40, Funong39, B73 (Maize), Z3116 (Sorghum), respectively. M, BM2000 + 1.5K DNA marker. The arrows show the expected PCR products/potential amplicons.

TABLE 4 Potential amplicon analysis results with PW2-23, PW2-28, and PW2-29 primer pairs in sugarcane (*S. spontaneum*), Maize (B73) and Sorghum (Z3116) genome.

Name	SSR motif	Potential amplicon with SSR motif			Potential amplicon without SSR motif		
		S. spontaneum	B73	Z3116	*S. spontaneum*	B73	Z3116
PW2-23	(CG)6	**265(5)***	**266(4)**	411(3)	None	160	120
		271(4)	443(3)			371	126
			522(3)			378	227
			858(3)			522	265
			1139(3)			641	371
			1315(3)			642	955
			1660(3)			647	1211
			1712(5)			690	1370
						837	1749
						1022	1769
						1633	1793
PW2-28	(GAG)5	**253(2)**	737(2)	None	None	209	None
		255(2)	738(2)				
		258(2)	738(2)				
			738(2)				
			738(2)				
			778(2)				
PW2-29	(CGT)5	**157(5)**	888(2)	1768(2)	None	60	131
		158(5)	1200(2)			60	335
		158(5)	1991(2)			328	744
							1961

*: 265(5) represents the size of potential amplicon is 265 bp and 5 SSR motif copies exist in amplicon, respectively. Bold means primer pair fully match with binding site.

EST-SSRs, 163 primer pairs proved effective for identifying 10 sugarcane accessions, demonstrating the suitability of transcriptome sequences as valuable resources for SSR markers' development.

The cluster results aligned well with the origin and pedigrees of 10 sugarcane accessions, providing insights into their relationships (Figure 6 and Table 1). For instance, sugarcane accessions from Fujian, Guangxi, and Taiwan clustered according to their breeding regions, while those with common parents clustered together. Additionally, our analysis revealed that *S. officinarum* shared a closer relationship with cultivated sugarcane compared to *S. spontaneum*. Interestingly, two cultivated sugarcane lines (Yunrui05-782 and Yunrui05-767) derived from hybrid wild species were distinct from other cultivated sugarcane varieties, highlighting the potential of wild species in expanding the genetic basis of cultivated sugarcane through sexual hybridization. We also explored the distribution of SSRs within the genomes of 10 sugarcane cultivars, observing a relatively high frequency of SSRs, approximately 1/4.43 kb. This frequency is comparable to certain other plant species such as *P. violascens* (1/4.45 kb), Chinese cabbage (1/4.67 kb), and Wheat (1/5.46 kb) but significantly higher than in Arabidopsis (1/13.83 kb) (Cardle et al., 2000; Peng and Lapitan, 2005; You et al., 2015; Cai et al., 2019). The types of repeat motifs in this study were not uniformly distributed in the sugarcane genome. In general, unlike former research studies on sugarcane (Table 4) by Singh et al. (2013); Xiao et al. (2020); Ukoskit et al. (2012), and Ul Haq et al. (2016), we found that the monomeric repeats accounted for the largest proportion, at 44.33% followed by tetrameric and dimeric repeats which were 39.68% and 13.10%, respectively (Table 3). These results are different from Xiao et al. (2020) in which trimeric repeats were most abundant. Dimeric and trimeric repeats were predominant when excluding monomeric repeats. Additionally, we found that the proportion of tetrameric, pentameric, and hexameric repeats was significantly lower than those reported by Xiao et al. [16] and other species (Table 5). Overall, our findings contribute to a deeper understanding of the SSR landscape in sugarcane and its implications for genetic studies and breeding programs.

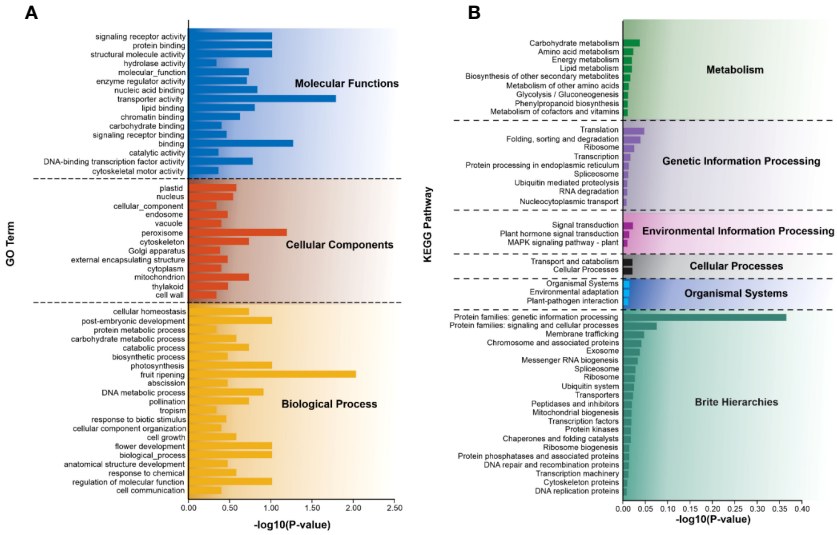


FIGURE 5
Summary of functional annotation of SSR-containing unigenes. **(A)** GO and **(B)** KEGG represent different classes based on the predicted function of the top 50 SSR-containing unigenes. The y-axis indicates the number of genes in each specific category.

TABLE 5 Comparison of frequency of microsatellites of different species.

Plant	Sugarcane*	Sugarcane**	Arabidopsis†	*Triticum aestivum*†	*Dendrocalamus latiflorus*†	*Phyllostachys violascens*†
Di-	23.55	22.06	26.27	20.77	16.1	48.06
Tri-	71.30	29.90	73.04	74.26	47.7	48.84
Tetra-	2.97	9.51	0.72	3.36	26.1	2.54
Penta-	1.05	24.03	0	1.12	6.9	0.42
Hexa-	1.15	15.48	0	0.5	3.3	0.14
Total	2761	37055	1070	43598	22305	9257

*this study; **Xiao et al., 2020; †Cai et al., 2019 all values in percent % except total number of SSR markers.

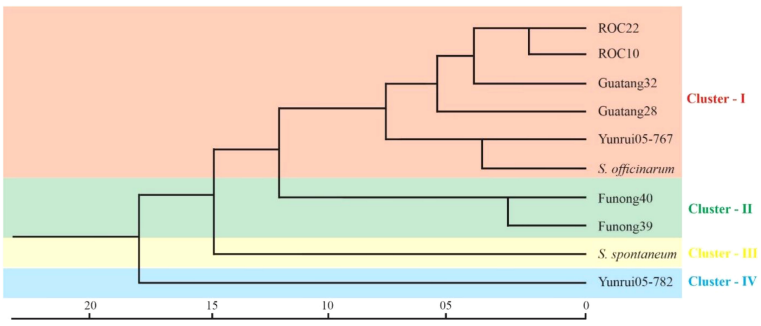


FIGURE 6
Cluster analysis of Sugarcane accession by SSR markers. The scale at the bottom represents the genetic distance between all the accessions.

As shown in Figure 2, the A/T motif was the predominant monomeric repeat (88%). In contrast, the GC/CT repeats accounted for 56%, which was higher than what was reported by Xiao et al. (2020) in sugarcane. Additionally, the abundance exceeded in other species such as taro (52.86%) (You et al., 2015), pigeon pea (16.7%) (Dutta et al., 2011), and wheat (8.7%) (Peng and Lapitan, 2005). Of trimeric repeats, CCG/CGG was the most predominant (48%), higher than the previous findings in taro (You et al., 2015),

sugarcane (4.84%) (Xiao et al., 2020), and rice and maize (Cardle et al., 2000). The CGC/GCG trimeric repeat at 17% was the second most abundant, which was lower than in *P. violascens* (3.45%) (Cai et al., 2019) and sugarcane (4.74%) (Xiao et al., 2020). The prevalence of trimeric repeat, CCG/CGG, a characteristic trimeric repeat in monocots was verified by our results but was rare in dicotyledonous plants (You et al., 2015; Cai et al., 2019; Xiao et al., 2020). The PIC is a critical metric in assessing the level of polymorphism of SSR markers, with a PIC value greater than 0.5 indicating a high level of polymorphism (Botstein et al., 1980). In our study, based on 163 EST-SSR markers, PIC values ranged from 0.292 to 0.972 with an average PIC value of 0.809 (Table S4). These findings align with Xiao et al. (2020) (0.70–0.94), Singh et al. (Singh et al., 2013) (0.12–0.99; 0.85), and Ul Haq et al. (2016) (0.51–0.93; 0.83).

In general, EST-SSR primer pairs and corresponding SSR loci were designed and aligned in *S. spontaneum*, sorghum, and maize in this study, which provided a possible way to develop EST-SSRs for sugarcane breeders. First, we developed EST-SSRs using sugarcane ES sequences or functional genes in the sugarcane genome. Some of the EST-SSR primer pairs were synthesized and successfully amplified by PCR in 10 sugarcane cultivars with sorghum and maize. Interestingly, our analysis revealed that a subset of SSR primer pairs (9 in *S. spontaneum*, 13 in maize, and 7 in sorghum) produced potential amplicons exclusively in one of these genomes. This observation suggests that while these species share some genetic similarities, they have also undergone unique evolutionary processes that have led to the development of distinct SSR loci. Such species-specific SSR markers can serve as important indicators of genetic divergence and could shed light on the evolutionary history of these species.

In sunflowers, most SSR-containing genes are involved in various biological processes such as cellular and metabolic processes (Lulin et al., 2012). Parmar et al., (Parmar et al., 2022) reported that most of the SRR-containing genes are involved in biological regulation and metabolic processes, which is consistent with the present study. The most important molecular functions of the GO-enriched genes in the present study are transport activity, binding, signaling receptor activity, protein activity, and catalytic activity. Additionally, the key biological processes associated with GO enrichment genes include fruit ripening, post-embryonic development, photosynthesis, and regulation of molecular functions. KEGG analysis of SSR-containing genes showed an important metabolic pathway such as carbohydrate metabolism and amino acid metabolism. The genetic information processing category was the second largest group.

5 Conclusion

In the present study, we achieved several significant outcomes. We successfully aligned sugarcane unigenes with sorghum and maize, leading to the identification and development of a valuable set of EST-SSR markers in sugarcane. A total of 4960 potential SSR markers were identified and of 240 randomly selected primer pairs, 173 were assessed for polymorphism. Among these, 163 primer pairs exhibited polymorphism when applied to 10 sugarcane accessions. Furthermore, we annotated 4203 SSR-containing unigenes into GO and KEGG databases, shedding light on their potential functions and pathways. Notably, we found that 56.43% of sugarcane unigenes mapped in maize genome to a single locus, 29.11% at two loci, 5.6% at three loci, and 8.58% with other loci. This suggests a distinct evolutionary relationship between sugarcane and sorghum with more duplication events occurring in maize chromosome segments. We believe these results have broad implications, contributing an important resource for future genomic and genetic studies in sugarcane but also serving as a powerful tool for studying evolutionary adaptation and genetic relationships in other related species.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

This article does not contain any studies with human participants or animals performed by any of the authors.

Author contributions

YL and PL: Conceptualization and Experimental Design, HJ: Data Collection, HJ and MW: Data Curation, YW and XZ: Resources—Plant materials Preparation, HJ and MW: drafted the manuscript, YL, SB, and PL: Review and Editing the drafted manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the initial funding of Guangxi University, grant number: XQZ130268.

Acknowledgments

The authors appreciate the help of Professor Fazhan Qiu (Huazhong Agricultural University) for providing us with the

genome-sequenced maize line (B73) and Vice Professor Guihua Zou (Institute of Nuclear Engineering, Zhejiang Academy of Agricultural Sciences) for providing the sorghum line (Z3116). We also appreciate Professor Ruiyuan Li for providing us with the Perl script to link the MISA and Primer3 programs. The authors appreciate the good suggestions from the editor and reviewers. The authors are grateful to Guangxi University for its financial support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1199210/full#supplementary-material

References

Biswas, M. K., Bagchi, M., Nath, U. K., Biswas, D., Natarajan, S., Jesse, D. M. I., et al. (2020). Transcriptome wide SSR discovery cross-taxa transferability and development of marker database for studying genetic diversity population structure of Lilium species. *Sci. Rep.* 10, 1–13. doi: 10.1038/s41598-020-75553-0

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* 32, 314.

Cai, K., Zhu, L., Zhang, K., Li, L., Zhao, Z., Zeng, W., et al. (2019). Development and characterization of EST-SSR markers from RNA-seq data in phyllostachys violascens. *Front. Plant Sci.* 10:50. doi: 10.3389/fpls.2019.00050

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156, 847–854. doi: 10.1093/genetics/156.2.847

Chen, J., Li, R., Xia, Y., Bai, G., Guo, P., Wang, Z., et al. (2017). Development of EST-SSR markers in flowering Chinese cabbage (Brassica campestris L. ssp. chinensis var. utilis Tsen et Lee) based on *de novo* transcriptomic assemblies. *PloS One* 12 (9), e0184736. doi: 10.1371/journal.pone.0184736

Cordeiro, G. M., Pan, Y.-B., and Henry, R. J. (2003). Sugarcane microsatellites for the assessment of genetic diversity in sugarcane germplasm. *Plant Sci.* 165, 181–189. doi: 10.1016/S0168-9452(03)00157-2

D'hont, A., Grivet, L., Feldmann, P., Glaszmann, J. C., Rao, S., and Berding, N. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Mol. Gen. Genet. MGG* 250, 405–413. doi: 10.1007/BF02174028

Dutta, S., Kumawat, G., Singh, B. P., Gupta, D. K., Singh, S., Dogra, V., et al. (2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [Cajanus cajan (L.) Millspaugh]. *BMC Plant Biol.* 11, 1–13. doi: 10.1186/1471-2229-11-17

Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K., et al (2018). A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nat. Commun.* 9(1):2638.

Gyawali, S., Harrington, M., Durkin, J., Horner, K., Parkin, I., Hegedus, D. D., et al. (2016). Microsatellite markers used for genome-wide association mapping of partial resistance to Sclerotinia sclerotiorum in a world collection of Brassica napus. *Mol. Breed.* 36, 1–13. doi: 10.1007/s11032-016-0496-5

Lulin, H., Xiao, Y., Pei, S., Wen, T., and Shangqin, H. (2012). The first Illumina-based *de novo* transcriptome sequencing and analysis of safflower flowers. *PloS One* 7, e38653. doi: 10.1371/journal.pone.0038653

Mccouch, S. R., Temnykh, S., Lukashova, A., Coburn, J., Declerck, G., Cartinhour, S., et al. (2001). Microsatellite markers in rice: abundance, diversity, and applications. *Rice Genet. IV. World Sci.*, 117–135. doi: 10.1142/9789812814296_0008

Oliveira, K. M., Pinto, L. R., Marconi, T. G., Mollinari, M., Ulian, E. C., Chabregas, S. M., et al. (2009). Characterization of new polymorphic functional markers for sugarcane. *Genome* 52, 191–209. doi: 10.1139/G08-105

Parmar, R., Seth, R., and Sharma, R. K. (2022). Genome-wide identification and characterization of functionally relevant microsatellite markers from transcription factor genes of Tea (Camellia sinensis (L.) O. Kuntze). *Sci. Rep.* 12, 201. doi: 10.1038/s41598-021-03848-x

Peng, J. H., and Lapitan, N. L. V. (2005). Characterization of EST-derived microsatellites in the wheat genome and development of eSSR markers. *Funct. Integr. Genomics* 5, 80–96. doi: 10.1007/s10142-004-0128-8

Singh, R. K., Jena, S. N., Khan, S., Yadav, S., Banarjee, N., Raghuvanshi, S., et al. (2013). Development, cross-species/genera transferability of novel EST-SSR markers and their utility in revealing population structure and genetic diversity in sugarcane. *Gene* 524, 309–329. doi: 10.1016/j.gene.2013.03.125

Sugita, T., Semi, Y., Sawada, H., Utoyama, Y., Hosomi, Y., Yoshimoto, E., et al. (2013). Development of simple sequence repeat markers and construction of a high-density linkage map of Capsicum annuum. *Mol. Breed.* 31, 909–920. doi: 10.1007/s11032-013-9844-x

Tóth, G., Gáspári, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967–981. doi: 10.1101/gr.10.7.967

Ukoskit, K., Posudsavang, G., Pongsiripat, N., Chatwachirawong, P., Klomsa-Ard, P., Poomipant, P., et al. (2019). Detection and validation of EST-SSR markers associated with sugar-related traits in sugarcane using linkage and association mapping. *Genomics* 111, 1–9. doi: 10.1016/j.ygeno.2018.03.019

Ukoskit, K., Thipmongkolcharoen, P., and Chatwachirawong, P. (2012). Novel expressed sequence tag-simple sequence repeats (EST-SSR) markers characterized by new bioinformatic criteria reveal high genetic similarity in sugarcane (*Saccharum* spp.) breeding lines. *Afr. J. Biotechnol.* 11, 1337–1363. doi: 10.1016/j.ygeno.2018.03.019

Ul Haq, S., Kumar, P., Singh, R. K., Verma, K. S., Bhatt, R., Sharma, M., et al. (2016). Assessment of functional EST-SSR markers (Sugarcane) in cross-species transferability, genetic diversity among poaceae plants, and bulk segregation analysis. *Genet. Res. Int.* 2016:7052323. doi: 10.1155/2016/7052323

Wang, Z., Fang, B., Chen, J., Zhang, X., Luo, Z., Huang, L., et al. (2010). *De novo* assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (Ipomoea batatas). *BMC Genomics* 11, 1–14. doi: 10.1186/1471-2164-11-726

Wu, J., Wang, Q., Xie, J., Pan, Y.-B., Zhou, F., Guo, Y., et al. (2019). SSR Marker-Assisted Management of Parental Germplasm in Sugarcane (*Saccharum* spp. hybrids) Breeding Programs. *Agronomy* 9 (8), 449. doi: 10.3390/agronomy9080449

Xiao, N., Wang, H., Yao, W., Zhang, M., Ming, R., and Zhang, J. (2020). Development and evaluation of SSR markers based on large scale full-length transcriptome sequencing in sugarcane. *Trop. Plant Biol.* 13, 343–352. doi: 10.1007/s12042-020-09260-5

You, Y., Liu, D., Liu, H., Zheng, X., Diao, Y., Huang, X., et al. (2015). Development and characterisation of EST-SSR markers by transcriptome sequencing in taro (Colocasia esculenta (L.) Schoot). *Mol. Breed.* 35, 134. doi: 10.1007/s11032-015-0307-4

Yu, B., and Zhang, C. (2011). "In silico PCR analysis," in *Methods in molecular biology* (Clifton, N.J.: Springer) 760, 91–107. doi: 10.1007/978-1-61779-176-5_6

Zalapa, J. E., Cuevas, H., Zhu, H., Steffan, S., Senalik, D., Zeldin, E., et al. (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* 99, 193–208. doi: 10.3732/ajb.1100394

Zhang, J., Nagai, C., Yu, Q., Pan, Y.-B., Ayala-Silva, T., Schnell, R. J., et al. (2012). Genome size variation in three Saccharum species. *Euphytica* 185, 511–519. doi: 10.1007/s10681-012-0664-6

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat. Genet.* 50 (11), 1565–73.

Zhou, Y., Wei, X., Abbas, F., Yu, Y., Yu, R., and Fan, Y. (2021). Genome-wide identification of simple sequence repeats and assessment of genetic diversity in Hedychium. *J. Appl. Res. Medicinal Aromatic Plants* 24, 100312. doi: 10.1016/j.jarmap.2021.100312

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 October 2023
DOI 10.3389/fpls.2023.1273740

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Juliano Lino Ferreira,
Embrapa Pecuária Sul, Brazil

*CORRESPONDENCE
Ashutosh Singh
singh.ashutosh026@gmail.com

RECEIVED 07 August 2023
ACCEPTED 09 October 2023
PUBLISHED 27 October 2023

CITATION
Divakar S, Jha RK, Kamat DN and Singh A (2023) Validation of candidate gene-based EST-SSR markers for sugar yield in sugarcane.
Front. Plant Sci. 14:1273740.
doi: 10.3389/fpls.2023.1273740

Validation of candidate gene-based EST-SSR markers for sugar yield in sugarcane

S. Divakar[1], Ratnesh Kumar Jha[2], D. N. Kamat[3] and Ashutosh Singh[2*]

[1]Department of AB&MB, CBSH, Dr. Rajendra Prasad Central Agricultural University (RPCAU), Samastipur, Bihar, India, [2]Centre for Advanced Studies on Climate Change, Dr. Rajendra Prasad Central Agricultural University (RPCAU), Samastipur, Bihar, India, [3]Sugarcane Research Institute, Dr. Rajendra Prasad Central Agricultural University (RPCAU), Samastipur, Bihar, India

Sugarcane (*Saccharum* spp.) is a widely cultivated crop that fulfils approximately 75% of the sucrose demand worldwide. Owing to its polyploidy and complex genetic nature, it is difficult to identify and map genes related to complex traits, such as sucrose content. However, association mapping is one of the alternatives for identifying genes or markers for marker-assisted selection. In the present study, EST-SSR primers were obtained from *in silico* studies. The functionality of each primer was tested using Blast2Go software, and 30 EST-SSR primers related to sugar content were selected. These markers were validated using association analysis. A total of 70 F1 diverse genotypes for sugar content were phenotypes with two check lines. All parameters related to sugar content were recorded. The results showed a significant variation between the genotypes for sugar yield traits such as Brix value, purity, and sucrose content, etc. Correlation studies revealed that the Brix%, sucrose content, and sucrose recovery were significantly correlated. An association analysis was performed using mixed linear model to avoid false positive associations. The association analysis revealed that the SEM 407 marker was significantly associated with Brix% and sucrose content. The SEM 407 primers are putatively related to diphosphate-fructose-6-phosphate 1-phosphotransferase which is associated with Brix% and sucrose content. This functional marker can be used for marker-assisted selection for sugar yield traits in sugarcane that could accelerate the sugarcane breeding program.

KEYWORDS

sugarcane, EST-SSR, candidate genes, sugar content, association mapping, marker-assisted selection

Introduction

Sugarcane (*Saccharum* spp.) belongs to the Poaceae/Gramineae family of the Andropogoneae tribe. *Saccharum* and its species were commercially used for sugar production owing to their high biomass and sucrose accumulation (Godshall and Legendre, 2003). The sugarcane genome is complex polyploid in nature, with *S. officinarum* having a basic chromosome number of $x = 10$ ($2n = 80$) and *S. spontaneum* $x = 8$ ($2n = 40$-128). Thus, there are two distinct chromosomes that coexist in modern cultivars (D'Hont et al., 1994; Zhang et al., 2018).

Sucrose is a commercial component of sugarcane, and the improvement of sugar recovery is the primary focus of any crop improvement program. The identification of genes or marker for sugar yield is an important strategy for the improvement of sugarcane. The mapping of genes is a promising tool for characterizing genetic architecture such as yield component traits, such as sucrose yield, cane yield, stalk diameter, stalk height, stalk number, and stalk weight, as well as resistance to diseases, pests, and abiotic stresses (Aitken et al., 2008; Welham et al., 2010; Singh et al., 2013; Gazaffi et al., 2014; Margarido et al., 2015; Balsalobre et al., 2016; Balsalobre et al., 2017; Yang et al., 2018). The complex polyploid and highly heterogeneous genetic nature of sugarcane association mapping could establish the QTL from linkage disequilibrium between the markers and the trait. Surveying a large number of genotypes in the existing germplasm of sugarcane can be helpful in finding associations between the markers and traits, using association mapping (Wei et al., 2006; Banerjee et al., 2015). To avoid spurious associations, the population structure and kinship of the association map population were employed to elucidate inferences (Lander and Schork, 2006). Validation of all those markers linked with QTL will have been identified by means of association mapping in a diverse population (Korir et al., 2013; Picañol et al., 2013; Ukoskit et al., 2019).

The expressed sequence tag (EST) database was used to identify the targeted SSR markers because ESTs are considered effective for the direct association with the trait of interest (Dudhe and Sarada, 2012). The interspecific transferability of expressed sequence tags derived from simple sequence repeats (EST-SSRs) and genomic SSRs is well established (Wen et al., 2010). EST-SSR primers are more beneficial than anonymous SSRs from untranslated regions (UTRs) or non-coding sequences, being frequently more transferrable between closely related genera (Pashley et al., 2006; Chapman et al., 2009). Due to the primer target sequences' location in the expressed DNA regions, which are predicted to be reasonably well preserved, there is a higher likelihood that the marker will be transferable across species borders (Varshney et al., 2005). EST-SSRs appear to disclose comparable amounts of polymorphism compared to SSRs found in UTRs despite their potential to reflect selectively harmful frame-shift mutations in coding areas. This is most likely because these coding regions have evolved to contain tri-nucleotide repeats (Ellis and Burke, 2007). Since EST-SSRs are physically connected to expressed genes, they constitute potentially useful markers. EST-SSR markers have a greater average rate of transferability between species than genomic SSRs because they reflect the expressed regions of a genome (Gupta et al., 2003). EST-SSR markers have been effectively used in gene tagging, linkage map construction, and QTL mapping (Qiu et al., 2010). Varietal crop improvement in various crops has become more feasible with the establishment of EST-SSR markers (Qiu et al., 2010; Ukoskit et al., 2019). EST-SSRs are highly regarded as a tool for breeding practices, perhaps because of their direct association with the genes of interest. It is also used in the identification of candidate genes in breeding and conservation input and population genetics studies (Yu et al., 2011). A mapping population obtained from a cross between commercial cultivars indulges in the introduction of EST-SSR markers into sugarcane linkage mapping (Oliveira et al., 2007; Palhares et al., 2012). There are some published reports of association mapping in sugarcane for traits such as biotic stress, cane yield, and sugar content (Wei et al., 2006; Wei et al., 2010; Débibakas et al., 2014; Banerjee et al., 2015; Gouy et al., 2015; Singh et al., 2016; Barreto et al., 2019; Fickett et al., 2019; Ukoskit et al., 2019; Coutinho et al., 2022). There are only a few studies that have investigated the identification of markers or genes for sugar yield traits using interspecific crosses (Reffay et al., 2005; Ukoskit et al., 2019). However, an association analysis requires a large population size and numerous EST-SSR primers. Moreover, these limitations could be avoided by choosing a diverse population with candidate genes for sugar yield for the validation of markers using an association study.

Therefore, the present study was conducted to identify different candidate gene-based EST-SSR markers from *in silico* studies, and these markers were validated using the association analysis of a diverse collection of sugarcane genotypes.

Materials and methods

Plant material

A total of 70 F1 diverse sugarcane genotypes for sugar yield traits were obtained from 14 different crosses, and five genotypes were chosen from each cross. The parents of all crosses were developed by crossing of *Saccharum officinarum* and *Saccharum spontaneum*. A few genotypes (BO102GC, BO137GC, and BO139GC) were also developed by general crosses (GC) for more variability. Seventy genotypes with two check lines (CoP16437 and CoP2061) were used in this study (Supplementary Table S2). These parents had contrasting natures for sugar and fiber yields. Because of limited seeds, these 70 genotypes were planted in an augmented complete block designed in the year 2021 at the Pusa farm of Dr. Rajendra Prasad Central Agricultural University, Samastipur, India. All genotypes were grown in seven different blocks, and each block had 10 genotypes with two check lines. The check lines were planted randomly in each block. The plot size of each block was 5.4 m^2. All standard agronomic practices were followed to raise the crops.

Phenotypic data and field data analysis

The population of 70 F1 genotypes was a phenotype for sugar-related traits after harvesting all genotypes. Brix, polarization (pol), sugar yield, and purity were recorded from four random stalks taken from each plot.

(1) Brix value (%): Brix value was measured using a hand-held refractometer. One degree of Brix is equal to the presence of 1 g sucrose in 100 g of the solution, and 1°Brix = 1% Brix. The strength of the solution was measured as a percentage of its mass. The reading was recorded using a refractometer with a sharp needle pierced through the stalk, and the collected substance was placed on a refractometer glass. The reading was recorded by the angle of the refractive index.

(2) Sucrose (%): Sucrose percentage was calculated using the following formula (Mehareb and Abazied, 2017):

$$\text{Sucrose (\%)} = \frac{\text{Brix \%} \times \text{purity \%}}{100}$$

(3) Purity (%): The percentage purity of the juice was calculated using the following formula (Mehareb and Abazied, 2017):

$$\text{Purity (\%)} = \frac{\text{Mass of the pure substance}}{\text{Mass of the impure sample}}$$

where pure sample: sucrose (%) and impure sample: Brix (%).

(4) Sugar recovery (CCS %): It was calculated using the following formula (Mehareb and Abazied, 2017):

$$\text{CCS \%} = [\text{Sucrose \%} - \{\text{Brix \%} - \text{Sucrose \%}\} \times 0.4] \times 0.73$$

(5) Sugar yield (t/ha)

Sugar yield was calculated using the following formula (Mehareb and Abazied, 2017):

$$\text{Sugar yield} = \frac{\text{Cane yield} \times \text{CCS \%}}{100}$$

Phenotypic data analysis

Pearson's correlation coefficients (r) between traits were calculated using the SAS CORR procedure based on trait means. Traits that were distributed normally with the Shapiro–Wilk test were considered normal data (Weber and Moorthy, 1952). Morphological data were used for the principal component analysis (PCA) using R studio. The data were imported in Excel format, and the eigenvalue must be greater than 1 for the variables. A bi-plot analysis was conducted using eigenvalues.

Extraction of DNA

A total of 500 mg of young leaf tissue was collected for marker analysis, and DNA was extracted using the cetyltrimethylammonium bromide method of Srivastava and Gupta (2008). DNA quantity and quality were determined using 1.0% agarose gel electrophoresis and nanodrop spectrophotometry, respectively.

Identification of a suitable EST-SSR marker and its validation by association study

Sugarcane is a polyploid crop, and its genome size and structure vary from genotype to genotype. Therefore, EST-SSR markers are best suited for tagging complex traits such as sugar yield. A total of 213 EST-SSR primers of sugarcane were identified in an *in silico* study (Supplementary Table S1). Furthermore, the functionality of these primers was tested using Blast2Go software. A total of 30 EST-SSR primers related to sugar yield were selected for this study based on their functionality (Supplementary Table S3). The PCR products were separated at 3.5–4.0% agarose gel. Out of 30, a total of 25 primers were amplified by PCR and used for further analysis. All 25 primers were scored based on the presence (1) or absence (0) of bands in the 70 genotypes of the mapping population.

Validation of EST-SSR primers using association mapping

The similarity coefficient among the genotypes was calculated using the genetic distance (Nei and Li, 1979). A neighbor-joining dendrogram was constructed using Past3 software. N-J analysis was performed using multivariate clustering, and the tree was constructed by Euclidean genetic distance with bootstrap replications of 100. The population structure was analyzed using STRUCTURE software to estimate the number of groups/subpopulations by setting the burning period length to 100,000, and each value of K was run three times with the K value varying from 1 to 10. Furthermore, a Q value below 0.9 was described as an admixture. An association analysis was performed using a mixed linear model as described by Yu and Buckler (2006). It was performed using TASSEL incorporating the Q matrix and K matrix to avoid false positives. The significant threshold for association was set at $P<0.05$.

Results and discussion

Sugarcane is polyploid crop with a complex genome, and it is difficult to interpret genome data. Less information about the sugarcane genome makes gene manipulation very difficult, but the identification of the functional genes responsible for sucrose accumulation makes it possible. EST-SSR markers were considered to be a highly regarded breeding tool as they are able to localize the functional gene by marker association (Palhares et al., 2012) since they may be directly associated with the gene expressing a particular trait.

Phenotypic yield data analysis

The germplasms used in this study were produced using an inter-varietal cross to validate primers for sugar yield traits in sugarcane. The yield distribution showed that the selected genotypes varied for sugar yield traits (Supplementary Figure S1). These germplasms were derived by crossing the sugarcane genotypes with contrasting sugar yield, and a few crosses were produced by general cross. The maximum, minimum, and mean values of all five phenotypic traits showed variation in the population for sugar yield traits. The Brix (%) values vary from

19.27% to 22.72%, with a mean of 20.65%. Similar trends were recorded for purity, sucrose content, sugar recovery, and sugar yield. This indicates that the selected germplasm shows variability and is appropriate for the study. The phenotypic correlations between the five traits were significant. The highest phenotypic correlation was found between sucrose content and sugar recovery, while the lowest phenotypic correlation was found between Brix value and purity (Table 1).

Population structure and genetic relationship

Many studies were conducted on sugarcane considering its complex polyploid genome, and several assumptions are not fulfilled for its complex structure; therefore, the applicability of this algorithm may be limited in sugarcane (Wei et al., 2006). Analyzing the sugarcane subpopulation using Structure software and mixed linear model provides an opportunity to track the gene related to complex traits of sugar yield in sugarcane. In the present study, minimum population size with a maximum variation was used, which is the most favorable for association analysis (Wei et al., 2006). Spurious associations were controlled, while the power to detect true associations was maximized using PCA as a random component to control for population structure (Pastina et al., 2012). PCA, as a random component, is included in the analysis, and the large population structure is captured with the first few axes that account for most of the variation, while the more subtle relationships among individuals are captured by the remaining significant axes. The population showed a clear and continuous variation in its structure, PCA, and neighbor-joining dendrogram (Fickett et al., 2019; Ukoskit et al., 2019). Furthermore, most of the structures found in these genotypes seem to originate from subtle kinship relationships rather than a large-scale population structure. The mapping population was diverse and highly heterozygous, and environmental conditions play a major role in the formation of sugar. The PCA results showed that Brix%, sucrose content, and CCS% were highly positively correlated, and the CCS and cane yield are highly negatively correlated (Figure 1). The results of the PCA and neighbor-joining dendrogram were coherent and showed no disjuncture in the population. The biplot of PCA overlays both individuals and the variables in a single graph. The loading range was varied from -5 to 5. The high absolute loadings were directed to either positively or negatively describe the variable that strongly influences the component, and a value less than that of the high loading indicates that they had a weak influence on the component. Sucrose content, Brix, and CCS% were highly positively correlated, as identified by PCA, and they showed 49% of the total variation in the phenotypic data (Figure 1). Hence, the traits directly related with sugar yield showed significant variability and indicate the diverse genotypic nature of germplasms.

The markers were identified by surveying the sugarcane database SUCEST, with their functionality scored by BLAST to determine the homology and putative function of the marker sequence (Supplementary Table S1). EST-SSR markers are derived from the expression regions of the genome and have greater potential for the direct association of the trait. The data Blast2Go showed that, at every 18.60 kb, one SSR motif was found to be very similar to cotton and wheat (Cardle et al., 2000). Pinto et al. (2004) reported the density of SSR in sugarcane at every 16.90 kb.

The neighbor-joining dendrogram is a bottom-up clustering method designed to provide a single tree and may be able to produce more than one dendrogram from the same data (Saitou and Nei, 1987). This method provides faster and better results than UPGMA, and most implementations provide a single tree (Page and Holmes, 2009). The dendrogram of the neighbor-joining relationship based on EST-SSR allele frequencies separated the populations into three differentiated clusters: A, B, and C. Cluster A has 33 genotypes, cluster B has 31 genotypes, and cluster C has six genotypes (Figure 2).

TABLE 1 Mean performance and the correlations between traits at the time of harvesting of sugarcane.

Statistics	Brix	Purity	Sucrose content	Sugar recovery	Sugar yield
Mean	20.65	85.88	18.12	14.33	10.4
Maximum	22.72	87.79	20.06	15.97	16.41
Minimum	19.27	83.45	16.9	13.2	6.76
Correlation[a]					
Brix					
Purity	0.127^{ns}				
Sucrose content	0.952*	0.139^{ns}			
Sugar recovery	0.947*	0.208*	0.981*		
Sugar yield	0.272*	0.345*	0.267*	0.301*	
Significant* *P<0.05*					

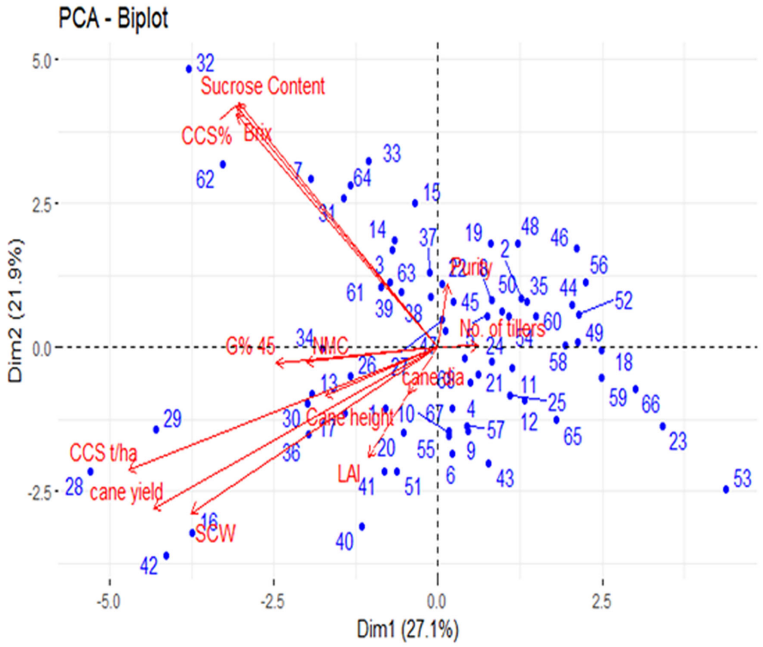


FIGURE 1
Biplot of principal component analysis PC1 (Dim1) and PC2 (Dim2); Where the dot represents the genotypes (70); Arrow represents the traits.

The Q matrix was calculated, and grouping of the population was performed using genotypic data. The result showed the marker-based kinship of the population. The number of subpopulations (ΔK) was identified by structure and maximized at $K = 5$. Populations with a single color were not shared by any other group, indicating that they were genetically distinct. The populations showed that the sharing of colors was similar to that of another group of individuals in the population (Figures 3, 4). The marker-based kinship matrix underlying the study of Yu et al. (2006) was determined based on the definition that random pairs of genotypes are unrelated, whereas Zhao et al. (2007) defined pairs of genotypes that do not share any allele as unrelated. The results clearly indicated that most of the genotypes were different.

Validation of EST-SSR markers

Sugarcane crops are polyploid and exhibit a high level of variation in the F1 generation. Therefore, validation of primers

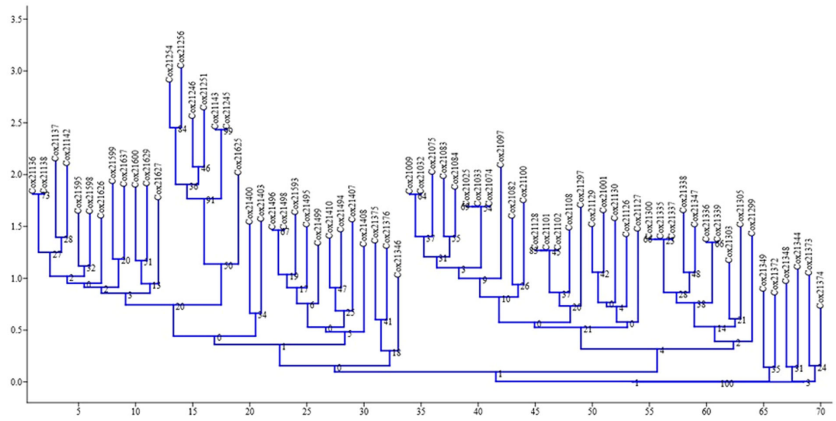


FIGURE 2
Dendrogram showing the relationship among 70 genotypes of sugarcane using EST-SSR markers.

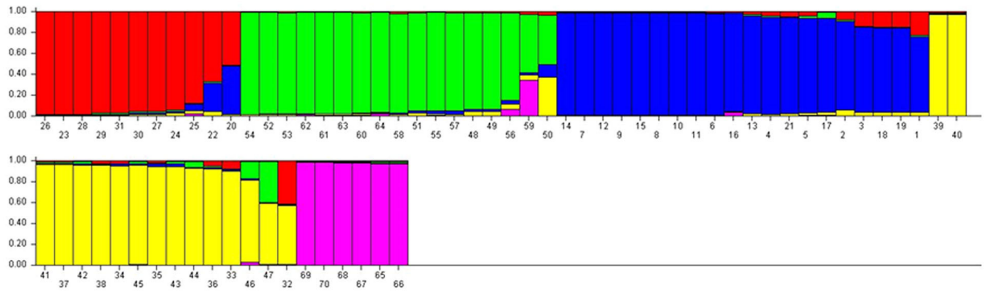


FIGURE 3
STRUCTURE analysis of bar plot. Populations with one solid color that is not shared by another group are genetically distinct Populations that share colors are more similar. Bar graphs for five sub-populations are indicated by different colors. The vertical coordinates indicate the membership coefficient of each individual and the horizontal represents the genotypes.

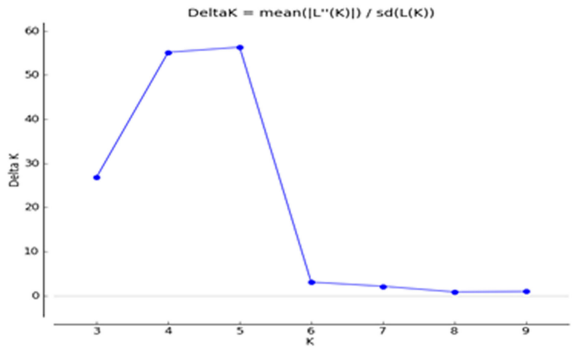


FIGURE 4
Delta K of STRUCTURE software using Evanno's criterion.

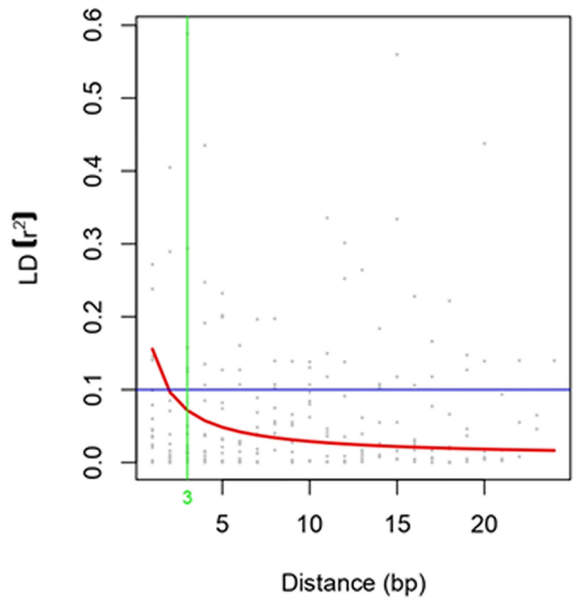


FIGURE 5
Linkage Disequilibrium Decay Plot.

using a simple chi-square is not possible. Hence, the primers were validated using association mapping. If the number of primers is a major limitation of the study, then the candidate gene approach for association mapping is best suited. A total of 30 EST-SSR markers were used for the association analysis of sugar yield traits. Of the 30 primers, 25 were amplified, and all these primers were tested for association studies. EST-SSR markers are a tool for association studies (Ukoskit et al., 2019; Coutinho et al., 2022). The EST-SSR markers related to sucrose content could be more effective than markers that focus on the varied functions of the gene. The linkage disequilibrium (LD) decay plot for the r^2 values between the markers was plotted against the genetic distance. The highest frequency of loci pair in LD is mapped less than 3 bp. The lowest frequency of loci pair was more than 20 bp, indicating that the probability of LD is low between distinct loci pairs. The majority of the loci pairs in LD with r^2 >0.01 at P<0.05 were found in ≤ 20bp (Figure 5). The values decreased as the genetic distance between the loci pairs increased. EST-SSR marker represented by the different regions of the genome was associated with the trait of interest at a P-value of 0.05. SEM407 was significantly associated with Brix%, sucrose content, and sugar recovery (CCS%) (Table 2). Except for SEM407, other EST-SSR markers did not show any association with sugar yield traits because the functional allele discovered in the mapping population might not be recognized in plants because it is rare in the larger germplasm. Compared to the genes tagged in the mapping population, the sugarcane accessions in the association population may have various trait-related alleles of various genes at various sites. Additional functional alleles that are absent from the mapping population may be found by validating the marker–trait relationship (Peace and Norelli, 2009). Although the number of markers used in this study is relatively low, the marker–trait association of SEM407 was significant. This marker was found to be significant for the sugar-related yield traits of genotypes. These results suggest that the association approach used in this study is consistent with the detection of QTL associated with sugar yield traits. The marker identified to the respective QTLs or genes should

TABLE 2 Associations study between EST-SSRs and sugar-related traits at P<0.05.

Markers	Association mappin2				
	Brix%	Purity	Sucrose content	Sugar recovery (CCS %)	Sugar yield
	P value	P value	P value	P value	P value
SEM2	0.07	0.47	0.08	0.1	0.2
SEM58	0.29	0.44	0.3	0.3	0.38
SEM112	0.3	0..2	0.29	0.3	0.2
SEM117	0.70	0.68	0.5	0.47	0.25
SEM159	0.70	0.77	0.64	0.62	0.22
SEM168	0.72	0.39	0.79	0.82	0.74
SEM191	0.99	0.26	0.84	0.90	0.56
SEM199	–	–	–	–	–
SEM203	0.43	0.91	0.41	0.43	0.96
SEM358	0.89	0.5	0.77	0.82	0.39
SEM368	0.12	0.85	0.3	0.27	0.48
SEM369	0.95	0.86	0.59	0.57	0.76
SEM407	0.002*	0.70	0.002*	0.002*	0.70
SEM425	0.29	0.40	0.2	0.1	0.55
SEM428	0.92	0.77	0.95	0.95	0.97
SEM430	0.58	0.61	0.56	0.58	0.38
SEM432	0.91	0.25	0.98	0.93	0.92
SEM433	0.63	0.46	0.7	0.63	0.56
SEM435	0.83	0.1	0.77	0.86	0.57
SEM436	0.73	0.87	0.62	0.63	0.41
SEM437	0.15	0.86	0.12	0.13	0.28
SEM439	0.76	0.12	0.74	0.80	0.11
SEM440	0.92	0.43	0.98	0.93	0.43
SEM454	0.46	0.96	0.59	0.56	0.30
SEM456	0.99	0.79	0.87	0.85	0.23

The experiment-wise threshold was based on the Bonferroni corrected method.
The thresholds for marker significance were at *P<0.05.

be used for marker-assisted selection (MAS). MAS for simply inherited traits are gaining increasing importance in breeding programs, allowing the acceleration of the breeding process of sugarcane (Francia et al., 2005). This study would be helpful for the plant breeders in marker-assisted selection in the prospect of achieving higher sugar yields while designing their crossing program.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

Author contributions

SD: Writing – original draft, Data curation, Formal Analysis, Methodology. RJ: Writing – review & editing, Funding acquisition, Supervision. DK: Writing – review & editing, Data curation. AS: Writing – original draft, Conceptualization, Funding acquisition, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Bihar government under Climate Resilient Agriculture Program (India).

Acknowledgments

The authors thank to Centre for Advanced Studies on Climate Change, RPCAU, Pusa and Department of Molecular Biology and Biotechnology, CBSH, RPCAU, Pusa Samastipur, Bihar, for providing the necessary infrastructure for us to carry out research work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1273740/full#supplementary-material

References

Aitken, K. S., Hermann, S., Karno, K., Bonnett, G. D., McIntyre, L. C., and Jackson, P. A. (2008). Genetic control of yield related stalk traits in sugarcane. *Theor. Appl. Genet.* 117, 1191–1203. doi: 10.1007/s00122-008-0856-6

Balsalobre, T. W. A., da Silva Pereira, G., Margarido, G. R. A., Gazaffi, R., Barreto, F. Z., Anoni, C. O., et al. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics* 18 (1), 1–19. doi: 10.1186/s12864-016-3383-x

Balsalobre, T. W., Mancini, M. C., Pereira, G. D. S., Anoni, C. O., Barreto, F. Z., Hoffmann, H. P., et al. (2016). Mixed modeling of yield components and brown rust resistance in sugarcane families. *Agro. J.* 108 (5), 1824–1837. doi: 10.2134/agronj2015.0430

Banerjee, N., Siraree, A., Yadav, S., Kumar, S., Singh, J., Kumar, S., et al. (2015). Marker-trait association study for sucrose and yield contributing traits in sugarcane (Saccharum spp. hybrid). *Euphytica* 205, 185–201. doi: 10.1007/s10681-015-1422-3

Barreto, F. Z., Rosa, J. R. B. F., Balsalobre, T. W. A., Pastina, M. M., Silva, R. R., Hoffmann, H. P., et al. (2019). A genome-wide association study identified loci for yield component traits in sugarcane (Saccharum spp.). *PloS One* 14 (7), e0219843. doi: 10.1371/journal.pone.0219843

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156 (2), 847–854. doi: 10.1093/genetics/156.2.847

Chapman, M. A., Hvala, J., Strever, J., Matvienko, M., Kozik, A., Michelmore, R. W., et al. (2009). Development, polymorphism, and cross-taxon utility of EST–SSR markers from safflower (Carthamus tinctorius L.). *Theor. Appl. Genet.* 120, 85–91. doi: 10.1007/s00122-009-1161-8

Coutinho, A. E., da Silva, M. F., Perecin, D., Carvalheiro, R., Xavier, M. A., de Andrade Landell, M. G., et al. (2022). Association mapping for sugarcane quality traits at three harvest times. *Sugar Tech.* 24 (2), 448–462. doi: 10.1007/s12355-021-01056-5

D'Hont, A., Lu, Y. H., León, D. G. D., Grivet, L., Feldmann, P., Lanaud, C., et al. (1994). A molecular approach to unraveling the genetics of sugarcane, a complex polyploid of the Andropogoneae tribe. *Genome* 37 (2), 222–230. doi: 10.1139/g94-031

Débibakas, S., Rocher, S., Garsmeur, O., Toubi, L., Roques, D., D'Hont, A., et al. (2014). Prospecting sugarcane resistance to sugarcane yellow leaf virus by genome-wide association. *Theor. Appl. Genet.* 127, 1719–1732. doi: 10.1007/s00122-014-2334-7

Dudhe, M. Y., and Sarada, C. (2012). Comparative assessment of microsatellite identification tools available in public domain. *DOR News Lett.* 18 (2), 8–9.

Ellis, J. R., and Burke, J. M. (2007). EST-SSRs as a resource for population genetic analyses. *Heredity* 99 (2), 25–132. doi: 10.1038/sj.hdy.6801001

Fickett, N., Gutierrez, A., Verma, M., Pontif, M., Hale, A., Kimbeng, C., et al. (2019). Genome-wide association mapping identifies markers associated with cane yield components and sucrose traits in the Louisiana sugarcane core collection. *Genomics* 111 (6), 1794–1801. doi: 10.1016/j.ygeno.2018.12.002

Francia, E., Tacconi, G., Crosatti, C., Barabaschi, D., Bulgarelli, D., Dall'Aglio, E., et al. (2005). Marker assisted selection in crop plants. *Plant Cell Tissue Organ Culture* 82, 317–342. doi: 10.1007/s11240-005-2387-z

Gazaffi, R., Margarido, G. R., Pastina, M. M., Mollinari, M., and Garcia, A. A. F. (2014). A model for quantitative trait loci mapping, linkage phase, and segregation pattern estimation for a full-sib progeny. *Tree Genet. Genomes* 10, 791–801. doi: 10.1007/s11295-013-0664-2

Godshall, M. A., and Legendre, B. L. (2003). *SUGAR| Sugarcane* Encyclopedia of Food Sciences and Nutrition (Second Edition), Elsevier.

Gouy, M., Rousselle, Y., Thong Chane, A., Anglade, A., Royaert, S., Nibouche, S., et al. (2015). Genome wide association mapping of agro-morphological and disease resistance traits in sugarcane. *Euphytica* 202, 269–284. doi: 10.1007/s10681-014-1294-y

Gupta, P. K., Rustgi, S., Sharma, S., Singh, R., Kumar, N., and Balyan, H. S. (2003). Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. *Mol. Genet. Genomics* 270, 315–323. doi: 10.1007/s00438-003-0921-4

Korir, N. K., Han, J., Shangguan, L., Wang, C., Kayesh, E., Zhang, Y., et al. (2013). Plant variety and cultivar identification: advances and prospects. *Crit. Rev. biotech.* 33 (2), 111–125. doi: 10.3109/07388551.2012.675314

Lander, E. S., and Schork, N. J. (2006). Genetic dissection of complex traits. *Focus* 265 (3), 2037–2458.

Margarido, G. R. A., Pastina, M. M., Souza, A. P., and Garcia, A. A. F. (2015). Multi-trait multi-environment quantitative trait loci mapping for a sugarcane commercial cross provides insights on the inheritance of important traits. *Mol. Breed.* 35, 1–15. doi: 10.1007/s11032-015-0366-6

Mehareb, E. M., and Abazied, S. R. (2017). Genetic variability of some promising sugarcane varieties (Saccharum spp) under harvesting ages for juice quality traits, cane and sugar yield. *Open Access J. Agric. Res.* 2 (2), 1–14. doi: 10.23880/OAJAR-16000127

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Nat. Aca. Sci.* 76 (10), 5269–5273. doi: 10.1073/pnas.76.10.5269

Oliveira, K. M., Pinto, L. R., Marconi, T. G., Margarido, G. R., Pastina, M. M., Teixeira, L. H. M., et al. (2007). Functional integrated genetic linkage map based on EST-markers for a sugarcane (Saccharum spp.) commercial cross. *Mol. Breed.* 20, 189–208. doi: 10.1007/s11032-007-9082-1

Page, R. D., and Holmes, E. C. (2009). *Molecular evolution: a phylogenetic approach*. (John Wiley & Sons).

Palhares, A. C., Rodrigues-Morais, T. B., Van Sluys, M. A., Domingues, D. S., Maccheroni, W., Jordão, H., et al. (2012). A novel linkage map of sugarcane with evidence for clustering of retrotransposon-based markers. *BMC Genet.* 13 (1), 1–16. doi: 10.1186/1471-2156-13-51

Pashley, C. H., Ellis, J. R., McCauley, D. E., and Burke, J. M. (2006). EST databases as a source for molecular markers: lessons from Helianthus. *J. Hered.* 97 (4), 381–388. doi: 10.1093/jhered/esl013

Pastina, M. M., Malosetti, M., Gazaffi, R., Mollinari, M., Margarido, G. R. A., Oliveira, K. M., et al. (2012). A mixed model QTL analysis for sugarcane multiple-harvest-location trial data. *Theor. Appl. Genet.* 124, 835–849. doi: 10.1007/s00122-011-1748-8

Peace, C., and Norelli, J. (2009). Genomics approaches to crop improvement in the Rosaceae. *Genet. Genomics Rosaceae*. 6, 19–53. doi: 10.1007/978-0-387-77491-6_2

Picañol, R., Eduardo, I., Aranzana, M. J., Howad, W., Batlle, I., Iglesias, I., et al. (2013). Combining linkage and association mapping to search for markers linked to the flat fruit character in peach. *Euphytica* 190, 279–288. doi: 10.1007/s10681-012-0844-4

Pinto, L. R., Oliveira, K. M., Ulian, E. C., Garcia, A. A. F., and De Souza, A. P. (2004). Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome* 47 (5), 795–804. doi: 10.1139/g04-055

Qiu, L., Yang, C., Tian, B., Yang, J. B., and Liu, A. (2010). Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (Ricinus communis L.). *BMC Plant Biol.* 10, 1–10. doi: 10.1186/1471-2229-10-278

Reffay, N., Jackson, P. A., Aitken, K. S., Hoarau, J. Y., D'Hont, A., Besse, P., et al. (2005). Characterisation of genome regions incorporated from an important wild

relative into Australian sugarcane. *Mol. Breed.* 15, 367–381. doi: 10.1007/s11032-004-7981-y

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4 (4), 406–425. doi: 10.1093/oxfordjournals.molbev.a040454

Singh, R. K., Banerjee, N., Khan, M. S., Yadav, S., Kumar, S., Duttamajumder, S. K., et al. (2016). Identification of putative candidate genes for red rot resistance in sugarcane (Saccharum species hybrid) using LD-based association mapping. *Mol. Genet. Genomics* 291, 1363–1377. doi: 10.1007/s00438-016-1190-3

Singh, R. K., Singh, S. P., Tiwari, D. K., Srivastava, S., Singh, S. B., Sharma, M. L., et al. (2013). Genetic mapping and QTL analysis for sugar yield-related traits in sugarcane. *Euphytica* 191, 333–353. doi: 10.1007/s10681-012-0841-7

Srivastava, S., and Gupta, P. S. (2008). Inter simple sequence repeat profile as a genetic marker system in sugarcane. *Sugar Tech.* 10, 48–52. doi: 10.1007/s12355-008-0008-y

Ukoskit, K., Posudsavang, G., Pongsiripat, N., Chatwachirawong, P., Klomsa-Ard, P., Poomipant, P., et al. (2019). Detection and validation of EST-SSR markers associated with sugar-related traits in sugarcane using linkage and association mapping. *Genomics* 111 (1), 1–9. doi: 10.1016/j.ygeno.2018.03.019

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genic microsatellite markers in plants: features and applications. *Trends Biotech.* 23 (1), 48–55. doi: 10.1016/j.tibtech.2004.11.005

Weber, C. R., and Moorthy, B. R. (1952). Heritable and nonheritable relationships and variability of oil content and agronomic characters in the F2 generation of soybean crosses 1. *Agr. J.* 44 (4), 202–209. doi: 10.2134/agronj1952.00021962004400040010x

Wei, X., Jackson, P. A., Hermann, S., Kilian, A., Heller-Uszynska, K., and Deomano, E. (2010). Simultaneously accounting for population structure, genotype by environment interaction, and spatial variation in marker–trait associations in sugarcane. *Genome* 53 (11), 973–981. doi: 10.1139/G10-050

Wei, X., Jackson, P. A., McIntyre, C. L., Aitken, K. S., and Croft, B. (2006). Associations between DNA markers and resistance to diseases in sugarcane and effects of population substructure. *Theor. Appl. Genet.* 114, 155–164. doi: 10.1007/s00122-006-0418-8

Welham, S. J., Gogel, B. J., Smith, A. B., Thompson, R., and Cullis, B. R. (2010). A comparison of analysis methods for late-stage variety evaluation trials. *Aust. New Z. J. Stat* 52 (2), 125–149. doi: 10.1111/j.1467-842X.2010.00570.x

Wen, M., Wang, H., Xia, Z., Zou, M., Lu, C., and Wang, W. (2010). Developmenrt of EST-SSR and genomic-SSR markers to assess genetic diversity in Jatropha Curcas L. *BMC Res. Notes* 3 (1), 1–8. doi: 10.1186/1756-0500-3-42

Yang, X., Islam, M. S., Sood, S., Maya, S., Hanson, E. A., Comstock, J., et al. (2018). Identifying quantitative trait loci (QTLs) and developing diagnostic markers linked to orange rust resistance in sugarcane (Saccharum spp.). *Front. Plant Sci. 9* 350. doi: 10.3389/fpls.2018.00350

Yu, J., and Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Curr. Opin. biotech.* 17 (2), 155–160. doi: 10.1016/j.copbio.2006.02.003

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38 (2), 203–208. doi: 10.1038/ng1702

Yu, Y., Yuan, D., Liang, S., Li, X., Wang, X., Lin, Z., et al. (2011). Genome structure of cotton revealed by a genome-wide SSR genetic map constructed from a BC1 population between Gossypium hirsutum and G. barbadense. *BMC Genomics* 12 (1), 1–14. doi: 10.1186/1471-2164-12-15

Zhang, J., Zhang, X., Tang, H., Zhang, Q., Hua, X., Ma, X., et al. (2018). Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. *Nat. Genet.* 50 (11), 1565–1573. doi: 10.1038/s41588-018-0237-2

Zhao, K., Aranzana, M. J., Kim, S., Lister, C., Shindo, C., Tang, C., et al. (2007). An Arabidopsis example of association mapping in structured samples. *PloS Genet.* 3 (1), e4. doi: 10.1371/journal.pgen.0030004

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 November 2023
DOI 10.3389/fpls.2023.1213476

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Dengfeng Hong,
Huazhong Agricultural University, China
Ning Zhao,
Anhui Agricultural University, China
Maolong Hu,
Jiangsu Academy of Agricultural Sciences (JAAS), China

*CORRESPONDENCE
Xiyuan Ni
nixiyuan@yeah.net
Tao Zheng
zhengtao@zaas.ac.cn

RECEIVED 28 April 2023
ACCEPTED 07 November 2023
PUBLISHED 23 November 2023

CITATION
Shi J, Yu H, Fu Y, Wang T, Zhang Y, Huang J, Li S, Zheng T, Ni X and Zhao J (2023) Development and validation of functional kompetitive allele-specific PCR markers for herbicide resistance in *Brassica napus*. *Front. Plant Sci.* 14:1213476. doi: 10.3389/fpls.2023.1213476

Development and validation of functional kompetitive allele-specific PCR markers for herbicide resistance in *Brassica napus*

Jianghua Shi[1], Huasheng Yu[1], Ying Fu[1], Tanliu Wang[1], Yaofeng Zhang[1], Jixiang Huang[1], Sujuan Li[2], Tao Zheng[3]*, Xiyuan Ni[1]* and Jianyi Zhao[1]

[1]Institute of Crop and Nuclear Technology Utilization, Zhejiang Academy of Agricultural Science, Hangzhou, China, [2]Central Laboratory, State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products, Zhejiang Academy of Agricultural Science, Hangzhou, China, [3]Institute of Biotechnology, Zhejiang Academy of Agricultural Science, Hangzhou, China

Effective weed control in the field is essential for maintaining favorable growing conditions and rapeseed yields. Sulfonylurea herbicides are one kind of most widely used herbicides worldwide, which control weeds by inhibiting acetolactate synthase (ALS). Molecular markers have been designed from polymorphic sites within the sequences of ALS genes, aiding marker-assisted selection in breeding herbicide-resistant rapeseed cultivars. However, most of them are not breeder friendly and have relatively limited application due to higher costs and lower throughput in the breeding projects. The aims of this study were to develop high throughput kompetitive allele-specific PCR (KASP) assays for herbicide resistance. We first cloned and sequenced *BnALS1* and *BnALS3* genes from susceptible cultivars and resistant 5N (*als1als1/als3als3* double mutant). Sequence alignments of *BnALS1* and *BnALS3* genes for cultivars and 5N showed single nucleotide polymorphisms (SNPs) at positions 1676 and 1667 respectively. These two SNPs for *BnALS1* and *BnALS3* resulted in amino acid substitutions and were used to develop a KASP assay. These functional markers were validated in three distinct BC_1F_2 populations. The KASP assay developed in this study will be valuable for the high-throughput selection of elite materials with high herbicide resistance in rapeseed breeding programs.

KEYWORDS

KASP assay, herbicide resistance, SNPs, ALS genes, marker-assisted selection

Highlights

- Developed KASP assays for *BnALS* genes are high throughput, low-cost, and capable of screening for herbicide-resistant alleles for marker-assisted selection.

Introduction

Rapeseed (*Brassica napus* L., AACC) is one of the most important oil-producing crops worldwide, with an annual production of more than 28 million tons of vegetable oil globally (USDA ERS, 2021) and also provides important raw material for biofuel and other industrial products (Ohlrogge, 1994; Thelen and Ohlrogge, 2002). Weeds, especially the broad leaf cruciferous species, are well adapted to compete with rapeseed for sunlight, water, soil nutrients and physical space in the fields (Miki et al., 1990; Larue et al., 2019). Hence, weeds are a significant problem and greatly limit rapeseed yield. The development of herbicide-tolerant varieties is a high priority for varietal development and the most cost-effective tool to manage weeds (Tan et al., 2005; Green, 2014).

Acetolactate synthase (ALS) is the key enzyme for the biosynthesis of the branched chain amino acids, including valine, leucine, and isoleucine (Duggleby et al., 2008; Garcia et al., 2017). ALS has been proved to be the target site of several important herbicides, such as sulfonylurea (SU), imidazolinone (IMI), triazolopyrimidine (TP), pyrimidinyl-thiobenzoates (PTB) and sulfonyl- aminocarbonyl-triazolinone (SCT) (Yu and Powles, 2014). ALS harboring amino acid substitutions caused by gene editing or ethyl methane sulfonate (EMS) mutagenesis has been found to confer high resistance to sulfonylurea herbicides in crops including wheat, rapeseed and watermelon (Tian et al., 2018; Zhang et al., 2019; Guo et al., 2020). The genome information derived from *Brassica napus* cultivars Darmor-bzh and ZS11 shows that there are five copies in the *BnALS* gene family (Chalhoub et al., 2014; Sun et al., 2017). Of these, *BnALS1* and *BnALS3* are highly conserved, and constitutively expressed in all tissues (Wu et al., 2020). Thus, *BnALS1* and *BnALS3* are regarded to be essential *ALS* housekeeping genes and the ideal herbicide-resistance targets for genetic manipulation (Rutledge et al., 1991; Wu et al., 2020).

Single nucleotide polymorphism (SNP) is a kind of DNA polymorphism in a genome which results from a single nucleotide change in a DNA sequence (Drenkard et al., 2000; Vignal et al., 2002). Because amino acid substitution caused by single nucleotide mutation may change protein function to some extent, single nucleotide changes provide new insights into protein function (Henikoff and Comai, 2003). Specific single nucleotide change can alter protein function, which is closely related with agronomic traits, then was used as an important tool for crops genetic improvement (You et al., 2018; Zhang et al., 2018). Functional markers derived from polymorphic sites within genes causally affect phenotypic variation. Functional markers are superior to random DNA markers such as RFLPs, SSRs and AFLPs owing to complete linkage with trait locus alleles, and are considered to be more accurate and efficient for gene identification and marker-aided selection (Andersen and Lubberstedt, 2003; Varshney et al., 2005; Zhou et al., 2013; Li et al., 2022). Over the past few decades, allele-specific PCR (AS PCR) markers, cleaved amplified polymorphic sequences (CAPS) markers, derived CAPS (dCAPS) markers and loop-mediated isothermal amplification (LAMP) markers were developed in plants based on single nucleotide polymorphisms (Michaels and Amasino, 1998; Drenkard et al., 2000; Zhou et al., 2013; Pan et al., 2014; Guo et al., 2020; Wu et al., 2020; Wang et al., 2022). All these markers are used to detect and select interesting traits by differentiating between homozygous and heterozygous states of plants. However, these markers require fragments separation by electrophoresis and/or digestion with restriction enzyme after PCR amplification, making their application relatively limited due to higher costs and lower throughput.

The development of user-friendly tools and platforms makes the wide-scale use and application of SNP markers possible in breeding programs. The KASP (kompetitive allele-specific PCR) genotyping assay utilizes a unique form of competitive allele-specific PCR combined with a novel, homogeneous, fluorescence-based reporting system for the identification and measurement of genetic variation occurring at the nucleotide level to detect SNPs (He et al., 2014). With the advantages of being low-cost and high throughput for genotyping SNPs, the KASP technology has been extensively used in the fields of human, animal and plant genetics (He et al., 2014; Semagn et al., 2013).

In this study, we aimed to develop the KASP assays for high-throughput genotyping for herbicide resistance. The SNPs were identified on the basis of Sanger sequencing of cloned *BnALS1* and *BnALS3* genes from both the resistant 5N and susceptible cultivars of *B. napus*. Allele-specific assays were developed on the basis of SNPs at positions 1676 and 1667 bp from *BnALS1* and *BnALS3*, respectively. The practical utility of the developed KASP assays was established by validating these in three segregating backcross progeny populations varying for herbicide resistance.

Materials and methods

Plant material

Three elite semi-winter *B. napus* cultivars (namely ZY50, ZY51 and ZS72) in Zhejiang province of China, and a double mutant 5N (*als1als1/als3als3*) with herbicide resistance (Guo et al., 2020) were used in this study. The seeds were sown usually in late September or early October and harvested around late May in Yangdu, Haining, Zhejiang Province.

Development of segregation populations with herbicide-resistant 5N

To obtain herbicide-resistant rapeseed lines with good agronomic and quality traits, three backcross progenies (BC_1s)

TABLE 1 List of primer sequences used for KASP assays.

Gene	Allele	Primer	Sequence (5′-3′)
BnALS1	G/T	KASP-C-1676-COM	TGGCGAACCCTGATGCGATTGTTGTGGAT
		KASP-C-1676-HEX	GAAGGTCGGAGTCAACGGATTTAGCTTTGTAGAACCGATCTTCCA
		KASP-C-1676-FAM	GAAGGTGACCAAGTTCATGCTGCTTTGTAGAACCGATCTTCCC
BnALS3	G/T	KASP-A-1667-COM	TGGCGAACCCTGATGCGATTGTTGTGGAC
		KASP-A-1667-HEX	GAAGGTCGGAGTCAACGGATTTAGCTTTGTAGAACCGATCTTCCA
		KASP-A-1667-FAM	GAAGGTGACCAAGTTCATGCTGCTTTGTAGAACCGATCTTCCC

were developed from crosses of ZY50/5N//ZY50, ZY51/5N//ZY51 and ZS72/5N//ZS72. The heterozygous lines (*ALS1als1/ALS3als3*) from these three BC_1F_1 populations were then screened using newly developed KASP markers (Table 1; Figures 1A, B) and self-pollinated to produce three distinct BC_1F_2 populations for further analysis. The plants were cultivated in the experimental fields located in Yangdu, Haining, Zhejiang province.

Amplification and Sequence analysis of *BnALS1* and *BnALS3* Genes

Genomic DNA of young rapeseed leaves from each plant was extracted with a modified cetyltriethylammnonium bromide (CTAB) method (Shi et al., 2017). Full-length *BnALS1* (2228 bp) and *BnALS3* (2027 bp) genes were isolated and amplified separately

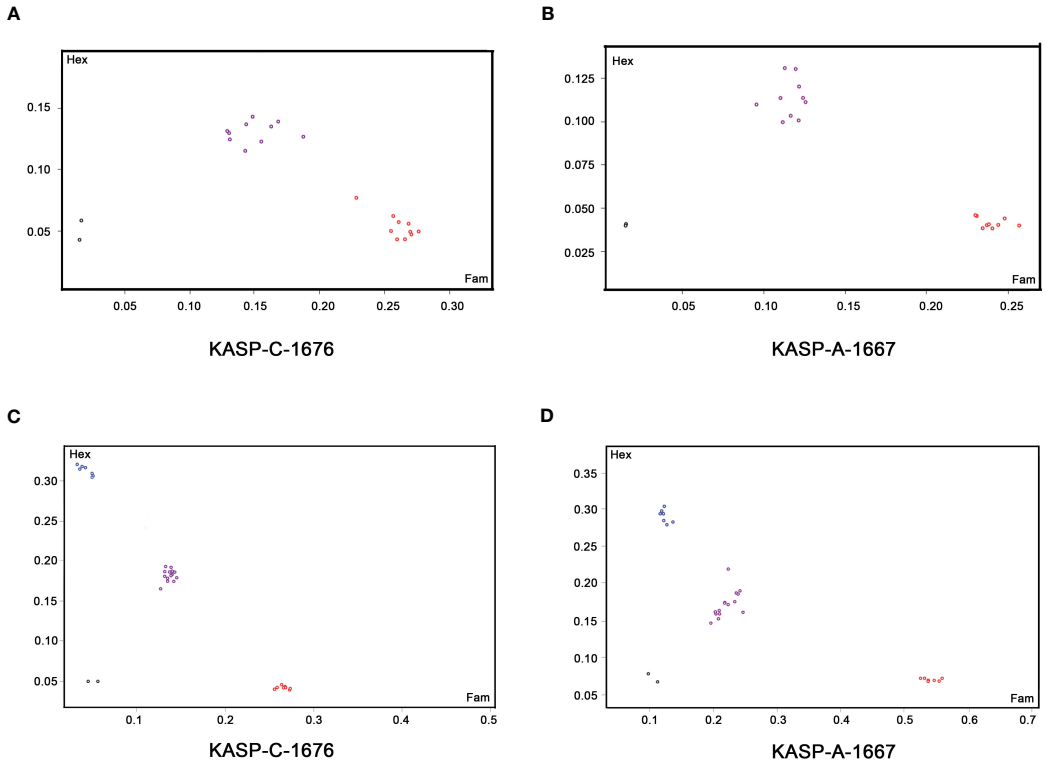


FIGURE 1
Kompetitive allele-specific PCR (KASP) genotyping using functional marker KASP-C-1676 and KASP-A-1667 in the segregating rapeseed populations for the selection of herbicide-resistant lines. **(A, B)** KASP genotyping of *BnALS1* gene on BC_1F_1 lines **(A)** and *BnALS3* gene on BC_1F_1 lines **(B)** respectively. The red, purple and black dots represent homozygous alleles (G/G), heterozygous alleles (G/T) and non-template control, respectively. **(C, D)** KASP genotyping of *BnALS1* gene on BC_1F_2 lines **(C)** and *BnALS3* gene on BC_1F_2 lines **(D)** respectively. The blue, red, purple and black dots represent homozygous alleles (T/T), homozygous alleles (G/G), heterozygous alleles (G/T) and non-template control, respectively.

from ZY50, ZY51, ZS72 and 5N using gene-specific primers as described (Guo et al., 2020). The resultant DNA fragments were sequenced by the Sanger dideoxy chain termination method on a capillary electrophoresis system (ABI 3730XL, Applied Biosystems, United States).

Nucleotide and amino acid multiple-sequence alignments were constructed using the CLUSTAL OMEGA program (Madeira et al., 2022) and colored by use of the GeneDoc 3.2 program with the default BLOSUM score. The sequence of the *ALS* gene from *A. thaliana* (GenBank accession no. NM_114714) was used as a reference. The nucleotide and amino acid sequences of *BnALS1* and *BnALS3* from three cultivars and 5N are listed in Supplementary data sheets 1, 2.

Primer design for *BnALS1* and *BnALS3* genes

BnALS1 and *BnALS3* sequences from susceptible cultivars (ZY50, ZY51 and ZS72) and the double mutant 5N were amplified and analyzed as mentioned above. Two herbicide-resistant SNPs, G1676T for *BnALS1* and G1667T for *BnALS3*, were used to develop functional markers. The KASP primers were designed according to the standard guidelines. Because *BnALS1* and *BnALS3* sequences are highly identical (97.6%), flanking sequence (including SNPs between *BnALS1* and *BnALS3*) of different alleles at each locus were extracted and used for primer design (Supplementary Figure 1). For each gene, the KASP marker consisted of two SNP-specific primers and one common primer. Of these three primers, two G/T alleles were linked to the FAM and HEX fluorescent linker-specific sequence of the LGC KASP reagents at the 5' end. The primer sequences are shown in Table 1.

Kompetitive allele-specific PCR genotyping

The genotyping assays of the developed KASP markers were performed on a 96-well plate. The KASP assay was performed in a 1.6 μL PCR reaction mix that consisted of 0.8 μL of KASP Master mix (LGC, Biosearch Technologies), approximately equal to 0.05 μL of primer, and 0.8 μL of DNA at a concentration of 10-20 ng/μL. The amplifications were performed using an IntelliQube (LGC, Biosearch Technologies) with the following cycling conditions: 94° C for 15min, 10 touchdown cycles (94°C for 20 s; touchdown at 61° C, dropping to -0.6°C per cycle 60 s) and followed by 26 cycles of amplification (94°C for 20 s, 55°C for 60 s).

Inheritance analysis

The susceptible rapeseed cultivars (ZY50, ZY51 and ZS72), 5N and the developed distinct BC_1F_2 populations were grown in the field, and seedlings at the 4-6 leaf stage were sprayed with tribenuron-methyl (TBM) at 20.25 g.a.i.ha^{-1}. Resistance of the parents and their derived BC_1F_2 populations was evaluated 20 days after treatment. The response phenotypes were scored as resistant (R) if they showed no herbicide damage or only slight injury, or susceptible (S) if they died. The segregation of each population was assessed using a Chi square test.

Further herbicide resistance analysis of the homozygous genotypes

Three distinct BC_1F_2 populations were derived from the crosses ZY50/5N//ZY50, ZY51/5N//ZY51 and ZS72/5N//ZS72. For each population, the seedlings of BC_1F_2 populations was analyzed for the four homozygous genotypes (AABB, AAbb, aaBB and aabb) using the composite KASP markers. These homozygous lines were then self-pollinated to generate BC_1F_3 seeds.

These BC_1F_3 homozygous lines from the three distinct BC_1F_2 populations were sown and grown in plastic pots (diameter, 10cm) containing a 1:1:1 mixture of peat moss, perlite and vermiculite under natural light conditions.

At least twenty BC_1F_3 seedlings from each of the four homozygous lines from the three distinct BC_1F_2 populations were sprayed with serial concentrations of 20.25, 30.38, 40.50 and 135 g.a.i.ha^{-1}TBM at the 4-6 leaf stage. Symptoms were recorded as resistant (R - no herbicide damage or only slight injury), mid-resistant (M - chlorosis or necrosis on some leaves, but no death) or susceptible (S - dead plants) at 20 days after the treatment.

Results

Phenotypic symptom of herbicide resistance

To observe the resistance to herbicide, the seedlings of ZY50, ZY51, ZS72 and 5N were sprayed with TBM at a concentration of 20.25 g.a.i.ha^{-1}. After exposure to TBM for 14 days, ZY50, ZY51 and ZS72 were growth injured with yellow or chlorotic leaves (Figures 2A, B). However, 5N, which harbored two resistant alleles, exhibited complete resistance, having no symptoms of chlorosis or necrosis (Figures 2A, B). Our results suggested that novel herbicide-resistant materials with good agronomic and quality traits could be developed through the crosses between the elite rapeseed cultivars and 5N.

Development of kompetitive allele-specific PCR marker for *BnALS1* and *BnALS3* genes

To detect single nucleotide polymorphisms (SNPs), *BnALS1* and *BnALS3* genes were separately cloned by PCR amplification from 5N and three cultivars, ZY50, ZY51 and ZS72 (Supplementary data sheet 1). Compared with 5N, these three cultivars have a common SNP at position G/T (1676) in *BnALS1* and a common SNP at position G/T (1667) in *BnALS3* (Supplementary Figure 1). A comparison of the amino acid sequences of susceptible cultivars/resistant 5N showed changes at W/L (474 in BnALS1; 471 in BnALS3) as compared to the changes at the two positions for the

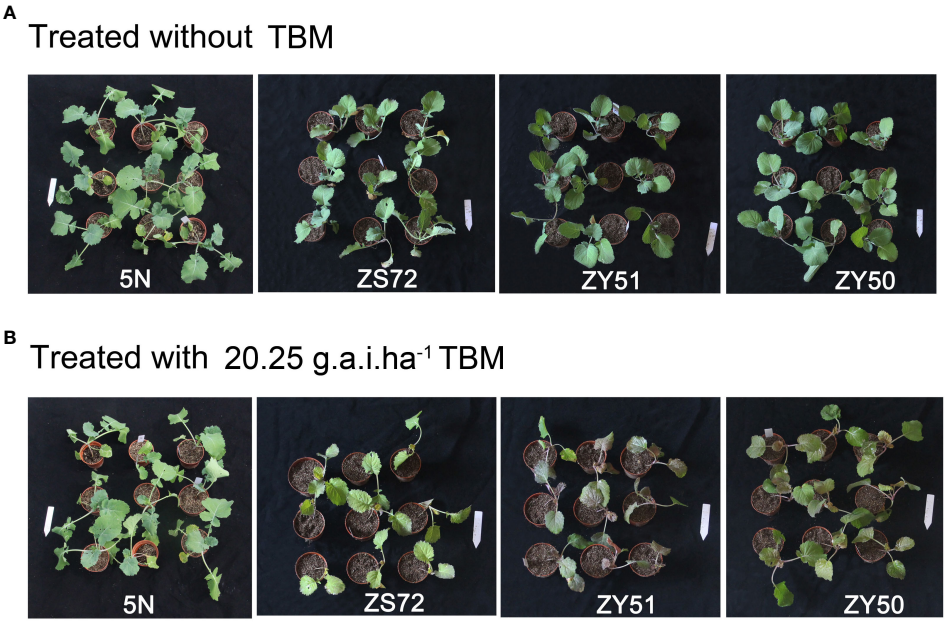


FIGURE 2
Herbicide resistance phenotypes of three cultivars ZY50, ZY51 and ZS72 and 5N mutant. **(A)** Phenotypes of cultivars ZY50, ZY51 and ZS72 and 5N mutant grown under normal condition. **(B)** Symptoms of cultivars ZY50, ZY51 and ZS72 and 5N mutant 14 days after treatment with 20.25 g.a.i.ha^{-1} TBM.

nucleotide sequence of *BnALS1* and *BnALS3* (Figure 3). A previous study proved that the substitutions of W/L in BnALS1 and BnALS3 could endow high herbicide resistance (Guo et al., 2020). Therefore, G1676T and G1667T were selected as the genotyping targets for *BnALS1* and *BnALS3* respectively. The KASP markers were designed for a SNP at position 1676 in *BnALS1* and for a SNP at position 1667 in *BnALS3* (Supplementary Figure 1). Both the marker KASP-C-1676 (specific to G1676T in *BnALS1*) and the marker KASP-A-1667 (specific to G1667 in *BnALS3*) could clearly distinguish type alleles GG, GT and TT among cultivars, cultivars/5N and 5N (Supplementary Figure 2). These two markers were also validated on BC_1F_1 populations, and formed separate clusters for heterozygous (GT) and homozygous (GG) alleles (Figures 1A, B).

Validation of kompetitive allele-specific PCR assays on BC_1F_2 populations

To confirm the KASP assay on herbicide resistance, three distinct BC_1F_2 populations were genotyped using KASP-C-1676 and KASP-A-1667. DNA was extracted from the first true leaves of seedlings before the herbicide treatment. Seedlings at the 4-6 leaf

FIGURE 3
Alignment of partial amino acid sequence of ALS proteins from Arabidopsis, ZY50, ZY51, ZS72 and herbicide-resistant mutant 5N (als1als1/als3als3). AtALS (GenBank accession no. NM_114714), BnALS1-ZY50, BnALS1-ZY51, BnALS1-ZS72, BnALS1-5N, BnALS3-ZY50, BnALS3-ZY51, BnALS3-ZS72 and BnALS3-5N (B. napus L. cv. ZY50, ZY51, ZS72 and 5N). The red arrow represents point mutations occurred in ALS1 and ALS3 of 5N. The shading of the alignment presents as follows: identical residues in black and different residues in dark gray. "*"indicates positions which have a single, fully conserved residue.

TABLE 2 Validation of the KASP assays for herbicide resistance in distinct BC_1F_2 populations of *B. napus*.

Genotype	BC_1F_2 (No of samples)			Phenotype
	ZY50/5N//ZY50	ZY51/5N//ZY51	ZS72/5N//ZS72	
AABB	5	5	6	S
AABb	24	24	20	R
AAbb	9	10	7	R
AaBB	16	16	20	R
AaBb	54	34	44	R
Aabb	22	22	27	R
aaBB	8	12	12	R
aaBb	18	17	24	R
aabb	4	6	9	R
Total	160	146	169	–
P value (1:2:1:2:4:2:1:2:1)	0.0958	0.5322	0.6678	–
$\chi 2$	13.5	7.0411	5.8166	–
–	0.0875	0.0875	0.0770	*P* value (1:15)
–	2.9192	2.9192	3.1277	$\chi 2$

A/a represents herbicide susceptible/resistant allele for BnALS1 gene; B/b represents herbicide susceptible/resistant allele for BnALS3 gene. The seedlings were treated with 20.25 g.a.i.ha^{-1} TBM at the 4-6 leaf stage, and phenotypes of the seedlings were observed 20 days after the treatment. R represents resistant to TBM; S represents susceptible to TBM.

stage were sprayed with TBM at a concentration of 20.25 g.a.i.ha^{-1}. Phenotypic symptoms were observed at 20 days after the treatment. The KASP assays were used for specific amplification of *BnALS1* and *BnALS3*. The frequency of the KASP alleles showed equivalence with the segregation expected for the BC_1F_2 populations (Figures 1C, D).

The combination of the KASP markers resulted in nine genotypes as shown by analysis of the seedlings in the BC_1F_2 populations (Table 2). These were AABB, AABb, AAbb, AaBB, AaBb, Aabb, aaBB, aaBb and aabb, and the ratio of isolation of these genotypes is 1:2:1:2:4:2:1:2:1 in the three distinct BC_1F_2 populations. In the three BC_1F_2 populations developed from crosses ZY50/5N//ZY50, ZY51/5N//ZY51 and ZS72/5N//ZS72, five, five and six homozygous plants with genotype AABB exhibited sensitivity to TBM treatment at 20 days after the treatment (Table 2). However, plants with other genotypes showed resistance to 20.25 g.a.i.ha^{-1} TBM treatment (Table 2). The ratio of susceptible lines to resistant lines is 1:15 in the three distinct BC_1F_2 populations (Table 2).

Further validation of genotype effect on herbicide resistance

To further validate the effect of genotype on herbicide resistance, we chose the BC_1F_3 plants with genotypes AABB, AAbb, aaBB and aabb developed from three distinct BC_1F_2 populations for resistance analysis. Plants with genotype AABB displayed serious damage with yellow leaves and eventual death within 20 days after treatment at all concentrations of TBM (Table 3). Plants with genotype AAbb exhibited resistance to 20.25-40.50 g.a.i.ha^{-1} TBM (Table 3). Plants with genotype aaBB showed resistance to 20.25-30.38 g.a.i.ha^{-1} TBM and mid-resistance

TABLE 3 The effects of four homozygous genotypes on herbicide resistance.

Genotype	TBM (g.a.i.ha^{-1})			
	20.25	30.38	40.50	135
AABB	S	S	S	S
AAbb	R	R	R	S
aaBB	R	R	M	S
aabb	R	R	R	R

A/a represents herbicide susceptible/resistant allele for BnALS1 gene; B/b represents herbicide susceptible/resistant allele for BnALS3 gene. R represents resistant to TBM;. M represents mid-resistant to TBM; S represents susceptible to TBM.

to 40.50 g.a.i.ha^{-1} TBM (Table 3). However, at higher concentration of 135 g.a.i.ha^{-1} TBM, Plants with genotype AAbb and aaBB showed chlorotic stunting, destroyed apex and eventual death (Table 3). By contrast to plants with genotype AAbb and aaBB, plants with genotype aabb exhibited complete resistance, having no chlorosis or necrosis, even to the higher concentration of 135 g.a.i.ha^{-1} TBM (Table 3).

Discussion

Successful weed management helps to improve crop yield in modern agricultural production systems. Resistant cultivars are the most effective and environmentally responsible strategy for protecting crops from weeds. Thus, developing new cultivars with high resistance to herbicides is now a major breeding objective in rapeseed. Acetolactate synthase encoded by *ALS* gene is responsible for biosynthesis of the branched chain amino acids, including valine, leucine, and isoleucine (Haughn and Somerville, 1990). The mutation of *ALS* gene may result in amino acid substitutions of ALS and inhibit the binding of the ALS enzyme with herbicides, which endows the plants with resistance to herbicide (Duggleby et al., 2008; Murphy and Tranel, 2019; Guo et al., 2020; Wu et al., 2020). Functional markers, such as AS-PCR (Hu et al., 2015) and CAPS (Li et al., 2015; Hu et al., 2017; Guo et al., 2020; Huang et al., 2020), have been developed to discriminate the allelic variation for the *ALS* genes. However, all these are gel based markers, and have relatively limited potential for high-throughput application. Thus, the development of a high-throughput and relatively cost-efficient marker system is important and necessary for improving breeding strategies.

As a key enzyme for the biosynthesis of branched chain amino acids, improper mutation of *ALS* can destroy its function and result in plant death. However, *ALS* harboring point mutations could confer sufficient tolerance to some kinds of herbicides with little damage to plant growth (Yu et al., 2010; Zhao et al., 2020; Cheng et al., 2021). We independently cloned and sequenced *BnALS1* and *BnALS3* from three cultivars ZY50, ZY51 and ZS72, and from the 5N mutant. DNA sequence alignment showed that 5N contains a single-nucleotide mutation (G1676T) in *BnALS1* and a single-nucleotide mutation (G1667T) in *BnALS3* based on sequence comparison with the three herbicide-susceptible cultivars; ZY50, ZY51 and ZS72 (Supplementary Figure 1), resulting in amino acid alterations, W474L (W574L, numbered according to ALS sequence in Arabidopsis) in BnALS1 and W471L (W574L) in BnALS3 (Figure 2). The W574L substitution has been reported to confer resistance to ALS inhibitors in rapeseed, sunflower and cocklebur (Bernasconi et al., 1995; Hattori et al., 1995; Sala et al., 2012; Hu et al., 2017; Guo et al., 2020). Mutation at P197 also conferred good tolerance to sulfonylureas in Arabidopsis, rapeseed and wheat (Li et al., 2015; Chen et al., 2017; Zhang et al., 2019; Huang et al., 2020; Wu et al., 2020; Cheng et al., 2021; Guo et al., 2022). In addition, mutations at the sites of Ala122, Ala205 and Ser653 of ALS have been reported to confer resistance to ALS inhibitors (Tan et al., 2005; Murphy and Tranel, 2019). These SNPs in *ALS* genes can be used for marker-assisted breeding.

5N is an important herbicide-resistant material with simultaneous mutations in *BnALS1* and *BnALS3* genes (Guo et al., 2020). We planned to design KASP markers for *BnALS1* and *BnALS3* genes in the 5N double mutant. Considering the highly similar (97.6%) sequence of *BnALS1* and *BnALS3*, it is difficult to develop high throughput markers capable of discriminating homozyous and heterozygous lines in these segregating populations. In this study, two KASP functional markers, KASP-C-1676 and KASP-A-1667, were successfully developed based on the specific characteristics of *BnALS1* and *BnALS3* genes from the cultivars and 5N (Table 1). These KASP markers will faciliate the use of 5N mutant for herbicide resistant rapeseed breeding.

The two KASP markers can clearly distinguish the genotypes of parents and hybrids (Supplementary Figure 2). Genotyping results performed by the two markers are highly consistent with the results of phenotypic evaluation (Figure 1; Table 2). Furthermore, these two KASP markers can distinguish the homozygous/heterozygous lines in three distinct segregated populations (BC_1F_2) developed from ZY50, ZY51, ZS72 and 5N (Table 2), which proved the high effectiveness of the KASP markers for genotyping under different genetic backgrounds. All these results suggested that the developed KASP markers are stable and effective to differentiate homozygous/heterozygous state of alleles in distinct populations and can be used for marker-assisted selection in rapeseed breeding projects.

In plants, synergistic effect is an important genetic phenomenon exhibited in the processes of hormone interaction, flower development and signal transduction (Poduska et al., 2003; Replogle et al., 2013; Yang et al., 2017). Synergistic effects have also been shown for herbicide resistance in crops. In *B. napus*, 5N (BnALS1-2R, W574L; BnALS3R, W574L) and DS3 (BnALS1-3R, P197L; BnALS3R, W574L) showed stronger herbicide resistance than mutants with single-point mutations (Guo et al., 2020 and Guo et al., 2022). In soybean, *Als1* (P197S) and *Als2* (W574L) exhibited synergistic resistant effects to ALS herbicides, and the combination of *Als1 and Als2* conferred stronger tolerance to SU (Walter et al., 2014). In this study, four homozygous genotypes were characterized and selected using the developed KASP markers. We analyze the effects of four genotypes on herbicide resistance. Our results showed that the lines containing two mutated alleles exhibited relatively stronger TBM resistance compared with those lines with a single mutated allele (Table 3), which is consistent with the findings reported previously (Guo et al., 2020). Altogether, these results suggested that the developed KASP markers are valuable functional markers and could be used for the high throughput selection of superior herbicide resistant materials by providing precise genotypic information, which will expedite the process of breeding herbicide-resistant rapeseed.

Conclusion

In this study, two KASP markers for *BnALS1* and *BnALS3*, KASP-C-1676 and KASP-A-1667, were successfully developed on the basis of SNPs in the *ALS* genes. These assays are highly gene specific and can effectively distinguish target genotype states. The developed KASP assays are high throughput and cost effective as

compared to gel-based markers and can be used for marker-assisted selection of herbicide resistance.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

JS conceived and coordinated the study. TW cloned and aligned the genes. TZ, YF, and SL conducted the KASP assays. XN, HY, YZ, and JH performed the field experiments and phenotypic data collection. JS wrote the manuscript and JZ revised it. All authors contributed to the article and approved the submitted version.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation of China (grant no. 31972875), the Project of Agriculture, Rural areas, Farmers and Nine Parties of Zhejiang Province (grant no. 2022SNJF010) and Key Laboratory of Digital Upland Crops of Zhejiang Province (grant no. 2022E10012).

Acknowledgments

We are grateful to Profs. Huiming Pu and Maolong Hu (Jiangsu Academy of Agricultural Science) for providing 5N material and to Dr. Yi Zhang (Molecular Marker (Wuhan) Biobreeding Co., LTD) for his assistance in marker designing. We are also grateful to Dr. Rebecca Horn at scientificproofreading.co.uk for her assistance in editing this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1213476/full#supplementary-material

References

Andersen, J. R., and Lubberstedt, T. (2003). Functional markers in plants. *Trends Plant Sci.* 8 (11), 554–560. doi: 10.1016/j.tplants.2003.09.010

Bernasconi, P., Woodworth, A. R., Rosen, B. A., Subramanian, M. V., and Siehl, D. L. (1995). A naturally occurring point mutation confers broad range tolerance to herbicides that target acetolactate synthase. *J. Biol. Chem.* 270 (29), 17381–17385. doi: 10.1074/jbc.270.29.17381

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A., Tang, H., Wang, X., et al. (2014). Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345 (6199), 950–953. doi: 10.1126/science.1253435

Chen, Y., Wang, Z., Ni, H., Xu, Y., Chen, Q., and Jiang, L. (2017). CRISPR/Cas9-mediated base-editing system efficiently generates gain-of-function mutations in Arabidopsis. *Sci. China Life Sci.* 60 (5), 520–523. doi: 10.1007/s11427-017-9021-5

Cheng, H., Hao, M., Ding, B., Mei, D., Wang, W., Wang, H., et al. (2021). Base editing with high efficiency in allotetraploid oilseed rape by A3A-PBE system. *Plant Biotechnol. J.* 19 (1), 87–97. doi: 10.1111/pbi.13444

Drenkard, E., Richter, B. G., Rozen, S., Stutius, L. M., Angell, N. A., Mindrinos, M., et al. (2000). A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. *Plant Physiol.* 124 (4), 1483–1492. doi: 10.1104/pp.124.4.1483

Duggleby, R. G., McCourt, J. A., and Guddat, L. W. (2008). Structure and mechanism of inhibition of plant acetohydroxyacid synthase. *Plant Physiol. Biochem.* 46 (3), 309–324. doi: 10.1016/j.plaphy.2007.12.004

Garcia, M. D., Wang, J. G., Lonhienne, T., and Guddat, L. W. (2017). Crystal structure of plant acetohydroxyacid synthase, the target for several commercial herbicides. *FEBS J.* 284 (13), 2037–2051. doi: 10.1111/febs.14102

Green, J. M. (2014). Current state of herbicides in herbicide-resistant crops. *Pest Manag Sci.* 70 (9), 1351–1357. doi: 10.1002/ps.3727

Guo, Y., Cheng, L., Long, W., Gao, J., Zhang, J., Chen, S., et al. (2020). Synergistic mutations of two rapeseed AHAS genes confer high resistance to sulfonylurea herbicides for weed control. *Theor. Appl. Genet.* 133 (10), 2811–2824. doi: 10.1007/s00122-020-03633-w

Guo, Y., Liu, C., Long, W., Gao, J., Zhang, J., Chen, S., et al. (2022). Development and molecular analysis of a novel acetohydroxyacid synthase rapeseed mutant with high resistance to sulfonylurea herbicides. *Crop J.* 10, 56–66. doi: 10.1016/j.cj.2021.05.006

Hattori, J., Brown, D., Mourad, G., Labbe, H., Ouellet, T., Sunohara, G., et al. (1995). An acetohydroxy acid synthase mutant reveals a single site involved in multiple herbicide resistance. *Mol. Gen. Genet.* 246 (4), 419–425. doi: 10.1007/BF00290445

Haughn, G. W., and Somerville, C. R. (1990). A mutation causing imidazolinone resistance maps to the csr1 locus of arabidopsis thaliana. *Plant Physiol.* 92 (4), 1081–1085. doi: 10.1104/pp.92.4.1081

He, C., Holme, J., and Anthony, J. (2014). SNP genotyping: the KASP assay. *Methods Mol. Biol.* 1145, 75–86. doi: 10.1007/978-1-4939-0446-4_7

Henikoff, S., and Comai, L. (2003). Single-nucleotide mutations for plant functional genomics. *Annu. Rev. Plant Biol.* 54, 375–401. doi: 10.1146/annurev.arplant.54.031902.135009

Hu, M., Pu, H., Gao, J., Long, W., Chen, F., Zhou, X., et al. (2017). Inheritance and molecular characterization of resistance to AHAS-inhibiting herbicides in rapeseed. *J. Integr. Agric.* 16, 2421–2433. doi: 10.1016/S2095-3119(17)61659-9

Hu, M., Pu, H., Kong, L., Gao, J., Long, W., Chen, S., et al. (2015). Molecular characterization and detection of a spontaneous mutation conferring imidazolinone

resistance in rapeseed and its application in hybrid rapeseed production. *Mol. Breed* 35, 46. doi: 10.1007/s11032-015-0227-3

Huang, Q., Lv, J., Sun, Y., Wang, H., Guo, Y., Qu, G., et al. (2020). Inheritance and molecular characterization of a novel mutated AHAS gene responsible for the resistance of AHAS-inhibiting herbicides in rapeseed (*Brassica napus* L.). *Int. J. Mol. Sci.* 21 (4), 1345–1362. doi: 10.3390/ijms21041345

Larue, C. T., Goley, M., Shi, L., Evdokimov, A. G., Sparks, O. C., Ellis, C., et al. (2019). Development of enzymes for robust aryloxyphenoxypropionate and synthetic auxin herbicide tolerance traits in maize and soybean crops. *Pest Manag Sci.* 75 (8), 2086–2094. doi: 10.1002/ps.5393

Li, H., Li, J., Zhao, B., Wang, J., Yi, L., Liu, C., et al. (2015). Generation and characterization of TBM herbicide-resistant rapeseed (Brasscia napus) for hybrid seed production using chemically induced male sterility. *Theor. Appl. Genet.* 128 (1), 107–118. doi: 10.1007/s00122-014-2415-7

Li, L., Sun, Z., Zhang, Y., Ke, H., Yang, J., Li, Z., et al. (2022). Development and utilization of functional kompetitive allele-specific PCR markers for key genes underpinning fiber length and strength in gossypium hirsutum L. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.853827

Madeira, F., Pearce, M., Tivey, A. R. N., Basutkar, P., Lee, J., Edbali, O., et al. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* 50 (W1), W276–W279. doi: 10.1093/nar/gkac240

Michaels, S. D., and Amasino, R. M. (1998). A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. *Plant J.* 14 (3), 381–385. doi: 10.1046/j.1365-313x.1998.00123.x

Miki, B. L., Labbe, H., Hattori, J., Ouellet, T., Gabard, J., Sunohara, G., et al. (1990). Transformation of *Brassica napus* canola cultivars with Arabidopsis thaliana acetohydroxyacid synthase genes and analysis of herbicide resistance. *Theor. Appl. Genet.* 80 (4), 449–458. doi: 10.1007/BF00226744

Murphy, B. P., and Tranel, P. J. (2019). Target-site mutations conferring herbicide resistance. *Plants (Basel)* 8 (10). doi: 10.3390/plants8100382

Ohlrogge, J. B. (1994). Design of new plant products: engineering of fatty acid metabolism. *Plant Physiol.* 104 (3), 821–826. doi: 10.1104/pp.104.3.821

Pan, L., Li, J., Zhang, W. N., and Dong, L. Y. (2014). Detection of the I1781L mutation in fenoxaprop-p-ethyl-resistant American sloughgrass (Beckmannia syzigachne Steud.), based on the loop-mediated isothermal amplification method. *Pest Manag Sci.* 71, 123–130. doi: 10.1002/ps.3777

Poduska, B., Humphrey, T., Redweik, A., and Grbic, V. (2003). The synergistic activation of FLOWERING LOCUS C by FRIGIDA and a new flowering gene AERIAL ROSETTE 1 underlies a novel morphology in Arabidopsis. *Genetics* 163 (4), 1457–1465. doi: 10.1093/genetics/163.4.1457

Replogle, A., Wang, J., Paolillo, V., Smeda, J., Kinoshita, A., Durbak, A., et al. (2013). Synergistic interaction of CLAVATA1, CLAVATA2, and RECEPTOR-LIKE PROTEIN KINASE 2 in cyst nematode parasitism of Arabidopsis. *Mol. Plant Microbe Interact.* 26 (1), 87–96. doi: 10.1094/MPMI-05-12-0118-FI

Rutledge, R. G., Quellet, T., Hattori, J., and Miki, B. L. (1991). Molecular characterization and genetic origin of the *Brassica napus* acetohydroxyacid synthase multigene family. *Mol. Gen. Genet.* 229 (1), 31–40. doi: 10.1007/BF00264210

Sala, C. A., Bulos, M., Altieri, E., and Weston, B. (2012). Response to imazapyr and dominance relationships of two imidazolinone-tolerant alleles at the Ahasl1 locus of sunflower. *Theor. Appl. Genet.* 124 (2), 385–396. doi: 10.1007/s00122-011-1713-6

Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2013). Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Mol. Breed* 33, 1–14. doi: 10.1007/s11032-013-9917-x

Shi, J., Lang, C., Wang, F., Wu, X., Liu, R., Zheng, T., et al. (2017). Depressed expression of FAE1 and FAD2 genes modifies fatty acid profiles and storage compounds accumulation in *Brassica napus* seeds. *Plant Sci.* 263, 177–182. doi: 10.1016/j.plantsci.2017.07.014

Sun, F., Fan, G., Hu, Q., Zhou, Y., Guan, M., Tong, C., et al. (2017). The high-quality genome of *Brassica napus* cultivar 'ZS11' reveals the introgression history in semi-winter morphotype. *Plant J.* 92 (3), 452–468. doi: 10.1111/tpj.13669

Tan, S., Evans, R. R., Dahmer, M. L., Singh, B. K., and Shaner, D. L. (2005). Imidazolinone-tolerant crops: history, current status and future. *Pest Manag Sci.* 61 (3), 246–257. doi: 10.1002/ps.993

Thelen, J. J., and Ohlrogge, J. B. (2002). Metabolic engineering of fatty acid biosynthesis in plants. *Metab. Eng.* 4 (1), 12–21. doi: 10.1006/mben.2001.0204

Tian, S., Jiang, L., Cui, X., Zhang, J., Guo, S., Li, M., et al. (2018). Engineering herbicide-resistant watermelon variety through CRISPR/Cas9-mediated base-editing. *Plant Cell Rep.* 37 (9), 1353–1356. doi: 10.1007/s00299-018-2299-0

USDA ERS. (2021). Available at: https://www.ers.usda.gov/data-products/oil-crops-yearbook/oil-crops-yearbook/.

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Sci.* 10 (12), 621–630. doi: 10.1016/j.tplants.2005.10.004

Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel Evol.* 34 (3), 275–305. doi: 10.1186/1297-9686-34-3-275

Walter, K. L., Strachan, S. D., Ferry, N. M., Albert, H. H., Castle, L. A., and Sebastian, S. A. (2014). Molecular and phenotypic characterization of Als1 and Als2 mutations conferring tolerance to acetolactate synthase herbicides in soybean. *Pest Manag Sci.* 70 (12), 1831–1839. doi: 10.1002/ps.3725

Wang, M. L., Zhi Tang, Z., Liao, M., Cao, H. Q., and Zhao, N. (2022). Loop-mediated isothermal amplification for detecting the Ile-2041-Asn mutation in fenoxaprop-P-ethyl-resistant Alopecurus aequalis. *Pest Manag Sci.* 79, 711–718. doi: 10.1002/ps.7239

Wu, J., Chen, C., Xian, G., Liu, D., Lin, L., Yin, S., et al. (2020). Engineering herbicide-resistant oilseed rape by CRISPR/Cas9-mediated cytosine base-editing. *Plant Biotechnol. J.* 18 (9), 1857–1859. doi: 10.1111/pbi.13368

Yang, Z. B., Liu, G., Liu, J., Zhang, B., Meng, W., Muller, B., et al. (2017). Synergistic action of auxin and cytokinin mediates aluminum-induced root growth inhibition in Arabidopsis. *EMBO Rep.* 18 (7), 1213–1230. doi: 10.15252/embr.201643806

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and applications of a high throughput genotyping tool for polyploid crops: single nucleotide polymorphism (SNP) array. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00104

Yu, Q., Han, H. P., Vila-Aiub, M. M., and Powles, S. B. (2010). AHAS herbicide resistance endowing mutations: effect on AHAS functionality and plant growth. *J. Exp. Bot.* 61 (14), 3925–3934. doi: 10.1093/jxb/erq205

Yu, Q., and Powles, S. B. (2014). Resistance to AHAS inhibitor herbicides: current understanding. *Pest Manag Sci.* 70 (9), 1340–1350. doi: 10.1002/ps.3710

Zhang, R., Liu, J., Chai, Z., Chen, S., Bai, Y., Zong, Y., et al. (2019). Generation of herbicide tolerance traits and a new selectable marker in wheat using base editing. *Nat. Plants* 5 (5), 480–485. doi: 10.1038/s41477-019-0405-0

Zhang, Y., Massel, K., Godwin, I. D., and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol.* 19 (1), 210. doi: 10.1186/s13059-018-1586-y

Zhao, N., Yan, Y. Y., Du, L., Zhang, X. L., Liu, W. T., and Wang, J. J. (2020). Unravelling the effect of two herbicide resistance mutations on acetolactate synthase kinetics and growth traits. *J. Exp. Bot.* 71 (12), 3535–3542. doi: 10.1093/jxb/eraa120

Zhou, L., Chen, Z., Lang, X., Du, B., Liu, K., Yang, G., et al. (2013). Development and validation of a PCR-based functional marker system for the brown planthopper resistance gene Bph14 in rice. *Breed Sci.* 63 (3), 347–352. doi: 10.1270/jsbbs.63.347

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 24 November 2023
DOI 10.3389/fpls.2023.1230559

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Ashish Prasad,
Kurukshetra University, India
Eui-Joon Kil,
Andong National University,
Republic of Korea

*CORRESPONDENCE
Ghulam Raza
graza4@gmail.com
Shahid Mansoor
shahidmansoor7@gmail.com

RECEIVED 29 May 2023
ACCEPTED 06 September 2023
PUBLISHED 24 November 2023

CITATION
Rahman SU, Raza G, Naqvi RZ, McCoy E, Hammad M, LaFayette P, Parrott WA, Amin I, Mukhtar Z, Gaafar A-RZ, Hodhod MS and Mansoor S (2023) A source of resistance against yellow mosaic disease in soybeans correlates with a novel mutation in a resistance gene.
Front. Plant Sci. 14:1230559.
doi: 10.3389/fpls.2023.1230559

A source of resistance against yellow mosaic disease in soybeans correlates with a novel mutation in a resistance gene

Saleem Ur Rahman[1,2], Ghulam Raza[1*], Rubab Zahra Naqvi[1], Evan McCoy[3], Muhammed Hammad[1], Peter LaFayette[3], Wayne Allen Parrott[3], Imran Amin[1], Zahid Mukhtar[1], Abdel-Rhman Z. Gaafar[4], Mohamed S. Hodhod[5] and Shahid Mansoor[1,6*]

[1]National Institute for Biotechnology & Genetic Engineering College, Pakistan Institute of Engineering and Applied Sciences (NIBGE-C, PIEAS), Faisalabad, Pakistan, [2]Department of Allied Health Sciences, Pak-Austria Fachhochscule-Institute of Applied Sciences and Technology (PAF-IAST), Mang, Haripur, Khyber Pakhtunkhwa, Pakistan, [3]Institute of Plant Breeding, Genetics & Genomics, University of Georgia, Athens, GA, United States, [4]Department of Botany and Microbiology, College of Science, King Saud University, Riyadh, Saudi Arabia, [5]Department of Biotechnology, October University for Modern Sciences and Arts, 6th of October City, Egypt, [6]Jamil ur Rehman Center for Genome Research, International Center for Chemical and Biological Sciences (ICCBS), University of Karachi, Karachi, Pakistan

Yellow mosaic disease (YMD) is one of the major devastating constraints to soybean production in Pakistan. In the present study, we report the identification of resistant soybean germplasm and a novel mutation linked with disease susceptibility. Diverse soybean germplasm were screened to identify YMD-resistant lines under natural field conditions during 2016-2020. The severity of YMD was recorded based on symptoms and was grouped according to the disease rating scale, which ranges from 0 to 5, and named as highly resistant (HR), moderately resistant (MR), resistant (R), susceptible (S), moderately susceptible (MS), and highly susceptible (HS), respectively. A HR plant named "NBG-SG Soybean" was identified, which showed stable resistance for 5 years (2016-2020) at the experimental field of the National Institute for Biotechnology and Genetic Engineering (NIBGE), Faisalabad, Pakistan, a location that is a hot spot area for virus infection. HS soybean germplasm were also identified as NBG-47 (PI628963), NBG-117 (PI548655), SPS-C1 (PI553045), SPS-C9 (PI639187), and cv. NARC-2021. The YMD adversely affected the yield and a significant difference was found in the potential yield of NBG-SG-soybean (3.46 $\pm$ 0.13^{a} t/ha) with HS soybean germplasm NARC-2021 (0.44 $\pm$ 0.01^{c} t/ha) and NBG-117 (1.12 $\pm$ 0.01^{d} t/ha), respectively. The YMD incidence was also measured each year (2016-2020) and data showed a significant difference in the percent disease incidence in the year 2016 and 2018 and a decrease after 2019 when resistant lines were planted. The resistance in NBG-SG soybean was further confirmed by testing for an already known mutation (SNP at 149th position) for YMD in the *Glyma.18G025100* gene of soybean. The susceptible soybean germplasm in the field was found positive for the said mutation. Moreover, an ortholog of the *CYR-1* viral resistance gene from black gram was identified in soybean as

Glyma.13G194500, which has a novel deletion (28bp/90bp) in the 5`UTR of susceptible germplasm. The characterized soybean lines from this study will assist in starting soybean breeding programs for YMD resistance. This is the first study regarding screening and molecular analysis of soybean germplasm for YMD resistance.

KEYWORDS

yellow mosaic disease, soybean, screening, resistance source, NBG-soybean

1 Introduction

Yellow mosaic disease (YMD) is one of the major devastating diseases that severely hampers soybean production. The disease is mainly caused by the mungbean yellow mosaic India virus (MYMIV), mungbean yellow mosaic virus (MYMV), horsegram yellow mosaic virus (HgYMV), and dolichos yellow mosaic virus (DoYMV). These viruses cause characteristic symptoms of yellow mosaic patterns on the leaf surfaces and are collectively named legume yellow mosaic viruses (LYMVs) (Qazi et al., 2007; Ilyas et al., 2009; Rahman et al., 2023a). Among LYMVs, HgYMV and DoYMV are rarely found, while MYMV and MYMIV are very common and infect many important legume crops (Maruthi et al., 2006; Rahman et al., 2023a). Previously, both MYMIV and MYMV species were found in India, while in Pakistan, MYMIV was the most frequent species found to infect major legume crops (Ilyas et al., 2009; Ilyas et al., 2010). However, recently a comprehensive study was conducted and found that MYMIV and MYMV have an equal role in causing infection in soybean cultivars in Pakistan (Rahman et al., 2023a; Rahman et al., 2023b). The exact data on yield loss due to YMD is not available, as the incidence of YMD varies in different locations and also varies for different crops (Varma and Malathi, 2003). In 1996, YMD caused a significant yield loss in soybean production, of approximately 105,000 metric tonnes. It has been reported that if the disease appears in the early stage of plant growth, the yield loss reaches up to 100% (Wrather et al., 1997; Kitsanachandee et al., 2013). Official reports on YMD from soybeans are lacking in Pakistan as soybean was only recently grown as a major legume in the country, however, anecdotal evidence from virologists suggests the disease is very common.

Identification of resistant soybean germplasm is a method of choice to prevent soybean cultivars from contracting YMD. In India, many resistant and susceptible soybean cultivars have been identified for YMD (Ram et al., 1984; Lal et al., 2005; Rani et al., 2017), but in Pakistan, no resistant cultivars have been identified or tested. Identification of virus-resistant germplasm is also the first step toward the identification of resistant (R) genes. These R genes have an important role in disease control. In 2012, the marker-assisted breeding technique was used for the identification of R genes and many genes have been identified that were found linked with YMD, such as the *CYR-1* gene in black gram (Maiti et al., 2012). The *CYR-1* gene was also found completely linked with MYMIV resistance in urdbean and mungbean (Maiti et al., 2011). In 2016, it was reported that the recessive form (*cyr-1*) of the *CYR-1* gene is a susceptibility factor for YMD in black gram (Patel et al., 2016). In urdbean, RGA-1, CEDG180, ISSR811, and YMV-1 were found to be closely linked with MYMIV resistance (Basak et al., 2005; Souframanien and Gopalakrishna, 2006; Gupta et al., 2013). In mungbean and black gram, the resistance gene analog (RGA) marker has been found to be linked with MYMV resistance. However, very little information is available on YMD-resistance genes in soybeans, except single nucleotide polymorphism (SNP) in an LRR-like protein kinase gene (chromosome 18; Glyma18g02850), which was found to be associated with soybean susceptibility to MYMIV (Yadav et al., 2015).

Studies regarding natural resistance sources for YMD and resistance genes from soybeans in Pakistan are lacking and the subject is in dire need of investigation. This study highlights the identification of both resistant and susceptible soybean germplasm and markers for YMD resistance in soybean cultivars, which will be used for the screening of resistant sources in soybean breeding programs in the future.

2 Materials and methods

2.1 Screening of soybean germplasm for YMD

A total of 1,007 soybean accessions were screened from 2016 to 2020. These accessions were acquired from the Plant Genetic Resources Institute (PGRI, Pakistan) and the Agriculture Research Service of the United States Department of Agriculture (USDA-ARS), USA. Local cultivars were provided by the National Agriculture Research Center (NARC), Islamabad, and Agriculture Research Center (ARS), Swat, Khyber Pakhtunkhwa, Pakistan as well as Ayub Agriculture Research Institute (AARI), Faisalabad, Pakistan. The imported germplasm, along with locally adapted cultivars, (Faisal soybean, NARC-2021, cv. Ajmeri) were grown in three replicates in single row plots that were 4.5 m long and were located at an experimental field, NIBGE, Faisalabad (31°25'0"N 73° 5'28" E). The trial location is a regional hub for begomoviruses of crops commonly grown in the area, such as cotton and mungbean (Habib et al., 2007; Ilyas et al., 2010). The plant-to-plant distance

was 10 cm and the row-to-row distance was 30 cm. The experiment was performed during the autumn seasons (from August to November) of the years 2016-2020. Recommended agronomic practices were performed for soybean management. The field was plowed two to three times before seed sowing. The recommended seed rate of 85 kg/ha of soybean was used. To achieve high nodulation and better nitrogen fixation, seeds were inoculated with plant growth-promoting rhizobacteria (PGPR) in the form of *Bradyrhizobium japonicum* (10 g per kg of soybean seeds). Recommended fertilizer rates of 25:60:50 kg/ha of Nitrogen (N), Phosphorus (P), and Potassium (K) were applied, respectively, by utilizing commercially available fertilizers (Urea, DAP, and SOP). During seedbed preparation, a full dose of P and K and a half dose of N were applied as basal doses. The remaining N was used at the flowering stage. To keep the crop free from weeds, the pre-emergence herbicide, Dual Gold (S-Metolachlor; Syngenta, Switzerland), at a concentration of 960g/L was used. Before sowing, the seeds were also treated with an antifungal, Hombre Ultra (Imidacloprid; Bayer, Germany). All the agronomic practices were kept uniform for all treatments except the soybean cultivars under study. The soybean crop was irrigated five to six times during the season. The field was routinely visited after sowing to document the appearance of symptoms of YMD. When YMD emerged, such as yellow mosaic spots on the leaf surface, the infection percentage was measured. The disease severity index (DSI)/percent infection (PI) of each cultivar was recorded. For DS, six different groups, namely, highly resistant (HR), resistant (R), moderately resistant (MR), moderately susceptible (MS), susceptible (S), and highly susceptible (HS), were formed as previously described (Islam et al., 2010) as shown in Table 1. HR represents high resistance to viruses having no symptoms. The percentage scale shows the plant infection severity of leaves and plant area affected. The DSI was calculated using the following formula (Habib et al., 2007):

$$individual\ \ DSI = \frac{\text{Sum of individual plant rating}}{\text{Total number of observed plants}} \times 100$$

Where DSI is the disease severity index, the disease rating is 0-5 as reported (Habib et al., 2007), and the individual plant rating was based on the symptoms of the leaves of each plant affected by the disease (Figure 1). The overall mean of disease rating for individual cultivars was recorded for the entire period (2016-2020). The percentage of disease incidence (PDI) was also measured by selecting the diseased (YMD) plants. In this case, the plants were classified as diseased or healthy irrespective of symptom severity. Several plants were randomly recorded at five positions in the entire field and classified as YMD or healthy. For each position, data from multiple plants were recorded. A total of 100 plants were classified in the entire field. The PDI was calculated using the following formula:

$$PI = \left[\frac{\text{Number of YMD plants}}{\text{Total number of observed plans}}\right] \times 100$$

2.2 Molecular characterization of viruses causing YMD

For the identification of begomoviruses causing YMD in soybeans, both symptomatic (susceptible) and asymptomatic (resistant) samples were collected, and the DNA was extracted using the modified CTAB method (Doyle and Doyle, 1987). The viral DNA was amplified using the primer pair MYMIV-F/MYMIV-R and MYMV-F/MYMV-R (Table S-2). A total of 10 PCR amplified products (MYMIV and MYMV) with respective sizes (~2.6-2.8 kb) were purified from agarose gel and cloned in a pTZ57R/T vector (Thermo Scientific, USA). The confirmed clones were sequenced using the Applied Biosystems 3730XL DNA sequencer. These clones were sequenced with M13 forward and M13 reverse primer pair, and then by primer walking strategy to get the complete sequence of each clone in both directions. Sequences were analyzed using Lasergene (DNAStar Inc.), and reads were assembled in SeqMan (Lasergene, DNAStar Inc., Madison, USA). After trimming the vector portion, a consensus contig was saved and analyzed. The sequences were submitted to the GenBank of NCBI (MN885463 and MK098184).

2.3 Confirmation of known SNP in soybean germplasm

For the confirmation of the already known resistant gene, *Glyma.18G025100* in soybean (Yadav et al., 2015), highly resistant and highly susceptible genotypes of soybean in Pakistan were selected. The total genomic DNA was extracted using a modified CTAB method (Doyle and Doyle, 1987) and the DNA was amplified using the primer pair: 18G025100-F: TCGTACTCA CGAAGGTGGA; 18G025100-R: AATGCGTTCTGAAGCTGTCC

TABLE 1 Criteria for percent infection of YMD in soybean germplasm.

Percent Infection	Disease Severity/Scale	Infection Category	Reaction Group
No symptoms	0	Highly resistant	**HR**
1-10%	1	Resistant	**R**
11-20%	2	Moderately resistant	**MR**
21-30%	3	Moderately susceptible	**MS**
31-50%	4	Susceptible	**S**
More than 50%	5	Highly susceptible	**HS**

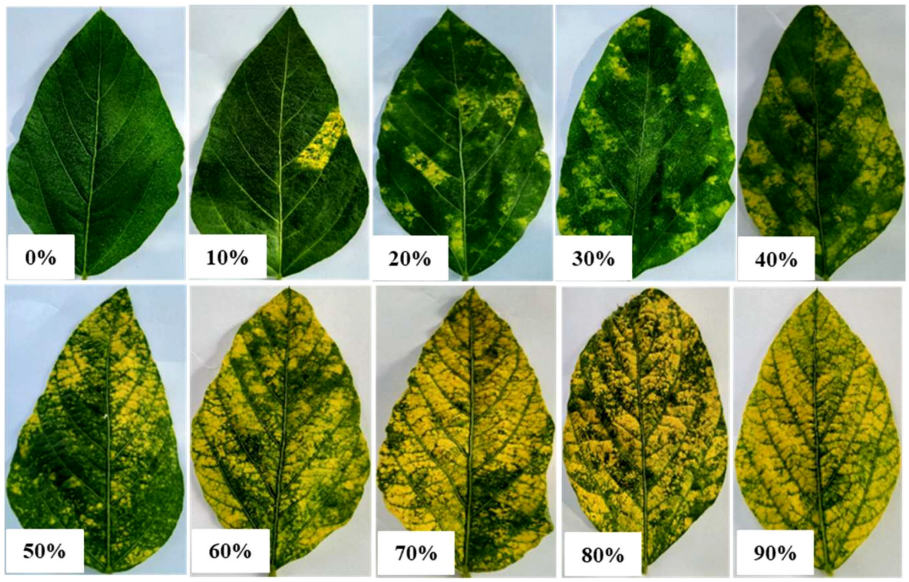


FIGURE 1
Representation of YMD severity percentage on soybean leaf.

(Yadav et al., 2015). The PCR products with the expected size (~2.6-2.8 kb) were gel-eluted and sequenced using the Applied Biosystems 3730XL DNA sequencer. The sequencing results were cleaned and compared using the BLASTn search tool in the NCBI data bank.

2.4 Confirmation of *CYR-1* ortholog in soybean

To identify the ortholog of the black gram *CYR-1* gene (Patel et al., 2016) in soybean, the complete sequence of the gene was retrieved from the Phytozome database (https://phytozome-next.jgi.doe.gov/), and BLASTP search engine was used to find the most similar sequences in the soybean genome. The most identical sequences of the predicted genes in soybeans were picked, which were further confirmed by wet lab experiments. The total DNA was extracted using a modified CTAB method (Doyle and Doyle, 1987) and amplified using a diverse range of overlapping primers (Table S-3) to sequence the complete gene. These overlapping primers were based on the predicted gene that followed the chromosome-walking strategy. These primers were designed in the available Geneious bioinformatics software. Overlapping primers were applied on both susceptible and resistant soybean cultivars identified in the present study. The PCR products were resolved using gel electrophoresis, purified and Sanger sequenced.

2.5 *In silico* analysis of 5`UTR and CYR-1 protein from resistant and susceptible germplasm

The transcription factor binding sites (TFBSs) were detected in both 5`UTR regions from the *CYR-1* gene from resistant and susceptible germplasm by using PlantPAN3.0 (http://plantpan.itps.ncku.edu.tw/). The ProtParam tool at Expasy (https://web.expasy.org/protparam/) was employed to identify the differences in CYR-1 protein from resistant and susceptible germplasm. The protein sequences of CYR-1 protein from resistant and susceptible germplasm were subjected to protein structure prediction and structure-based alignment by I-TASSER (https://zhanggroup.org/I-TASSER/) and TM-align (https://zhanggroup.org/TM-align/). Protein structures were visualized by PyMOL2.5 (https://pymol.org/2/). The interaction of both the CYR-1 gene from resistant germplasm and the CYR-1 gene from susceptible germplasm with the MYMIV viral coat protein was checked using PSOPIA (https://mizuguchilab.org/PSOPIA/) and ISLAND (https://island.pythonanywhere.com/).

2.6 Statistical analysis

The data were analyzed using a one-way analysis of variance (ANOVA) and the Tukey test was applied at α =0.05 (95% interval) using GraphPad Prism 6 (https://graphpad-prism.software.informer.com/6.0/).

3 Results

3.1 Evaluation of soybean germplasm for YMD

The soybean accessions (Table S-1) were evaluated during five successive years 2016-2020 in the autumn season (from August to November). In the year 2016, out of 128 entries, 18 entries found HR and 30 were HS (Table 2). The disease severity index was higher in the year 2016 (Figure 2). The HR and HS entries proceeded

10.3389/fpls.2023.1230559

TABLE 2 Number of YMD resistant and susceptible soybean genotypes during autumn (2016-2020).

Reaction	Number of Genotypes					
	2016	2017	2018	2019	2020	Total
HR	18	01	01	04	1	25
R	23	00	10	32	6	71
MR	11	00	32	56	11	110
MS	12	01	47	79	16	155
S	34	06	37	127	19	223
HS	30	28	174	160	31	423
Total	**128**	**36**	**301**	**458**	**84**	**1007**

further along with those accessions having high yield and resistance to YMD. There were more resistant lines in 2019 as crosses and mutants were included, which increased the resistance gene pool further. The individual leaf disease severity scale was converted to individual plants and identified HS and HR soybean germplasm (Figure 3). A highly resistant plant, selected from cv. Jack (PI540556) and named SG-soybean, was identified in the year 2016. The resistance in this line was stable in the entire growing period (2016-2020). The SG-soybean line showed complete resistance with no viral symptoms (Figure 4), while the approved cultivars, such as cv. Ajmeri and NARC-2021, were found to be HS. The accession NBG-117 (PI548655) and exotic genotypes were also found to be susceptible. The cv. NARC-2021 was previously recorded as NARC-16, which was approved recently in the year 2023 for general cultivation in Khyber Pakhtunkhwa, Pakistan. The incidence of disease in susceptible cultivars was more in check cultivars: cv. Ajmeri and NARC-2021 (Figure 4). The persistent nature of these cultivars to virus resistance showed that the susceptibility and resistance in these cultivars were stable (Table 3). The SG-soybean line was not only resistant to YMD but was resistant to multiple viruses. To have a complete picture of each cultivar, the reaction of each cultivar in each year and the overall mean reaction are summarized in Table 3. The YMD incidence was measured in the years 2016-2020 (Figure 2). The disease incidence was found to be very high (Figure 2) in 2016 and 2019, while it was significantly low ($P< 0.05$) in the years 2018 and 2020.

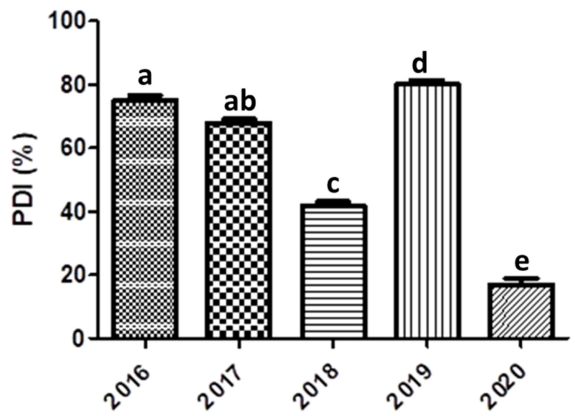


FIGURE 2
Year-wise YMD incidence (%) in soybeans. Lower case letters show significant difference.

3.2 Yield and yield-linked traits

The yield of the resistant SG-Soybean line was higher (3.46 ± 0.13a t/ha) than that of the susceptible NARC-2021 (0.44 ± 0.01c t/ha) and NBG-117 (1.12 ± 0.01d t/ha) (Table 3). The increase in SG soybean was ~7.9 times that of NARC-2021 and ~3 times that of NBG-117 (Table 3). Although the SPP and PPP of HS germplasm (NARC-2021 and NBG-117) were high, the germination growth was badly affected by YMD in these cultivars, so the number of plants in these germplasm was lower as compared to the SG-soybean line; hence the yield of these HS soybean germplasm was less than that of the SG-soybean line (Table 3).

3.3 Molecular analysis for YMD

As the YMD is caused by both MYMIV and MYMV, to confirm the YMD for both the viral strains, the susceptible and resistant cultivars were evaluated (Table 2). All the YMD symptomatic plants were positive for MYMIV/MYMV. The sequences of MYMIV and MYMV identified in the field were submitted in NCBI and are available under accession nos. MN885463 and MK098184, respectively.

3.4 Confirmation of known SNP in soybean resistant gene

In the current study, the highly resistant and susceptible germplasm of soybeans in Pakistan were checked for the known SNP in the *Glyma.18G025100* gene on chromosome 18. The primer pair: 18G025100-F: TCGTACTCACGAAGGTGGA and 18G025100-R: AATGCGTTCTGAAGCTGTCC was used for the amplification of the gene carrying the SNP for C to G transversion. The gel electrophoresis confirmed the expected fragment size

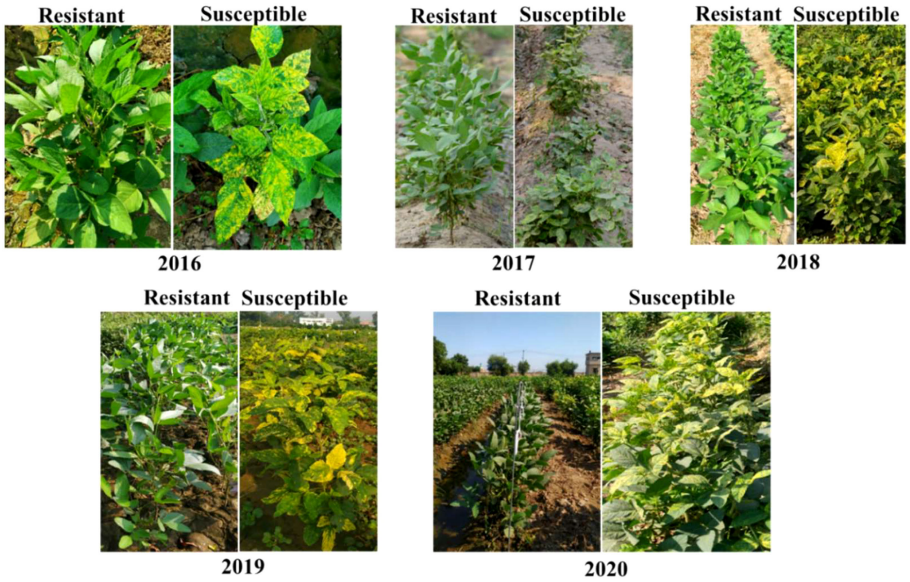


FIGURE 3
Resistant and susceptible soybean germplasm for YMD tested in 2016-2020.

FIGURE 4
Pictorial view of resistant and susceptible soybean germplasm under natural field conditions.

TABLE 3 Agronomic performance of highly resistant and susceptible soybeans.

Genotype	DF	DM	PH (cm)	PPP	SPP	GY (t/ha)
SG-soybean	41 ± 3.0^{a}	98 ± 2.0^{a}	37.3 ± 2.0^{a}	70.6 ± 3.0^{a}	2.86 ± 0.15^{a}	3.46 ± 0.13^{a}
NARC-2021	65 ± 2.6^{c}	91.6 ± 3.5^{a}	38.3 ± 7.2^{a}	93.6 ± 5.5^{b}	2.93 ± 0.05^{a}	0.44 ± 0.01^{c}
NBG-117	40 ± 2.5^{a}	102.3 ± 2.5^{ab}	62.6 ± 8.0^{ab}	127.3 ± 12.0^{bc}	2.96 ± 0.05^{a}	1.12 ± 0.01^{d}

The small letters "a-c" etc shows the significant difference.

(Figure 5A). However, the PCR product was further confirmed by Sanger sequencing, which found the same C to G transversion, as shown in Figure 5B.

3.5 Confirmation of the *CYR-1* gene ortholog in soybean

To identify the ortholog of the *CYR-1* gene in soybean, the sequence of the *CYR-1* gene was aligned with the soybean genome. The sequence of the *CYR-1* gene of black gram showed 65% identity with the soybean gene *Glyma.13G194500*. The overlapping primers were designed to cover the whole gene with 100 bp apart from the gene covering 5` UTR and 3` UTR (Table S-3; Figure 6A). The 128 amino acid deletion was not identified in the closely related gene, however, in one susceptible variety (Minsoy), a deletion (28bp) in the 5` UTR was identified. This deletion was also observed when resolved on agarose gel (Figure 6B). The deletion was confirmed by Sanger sequencing, which was identified in 5` UTR of *Glyma.13G194500* (Figure 6C).

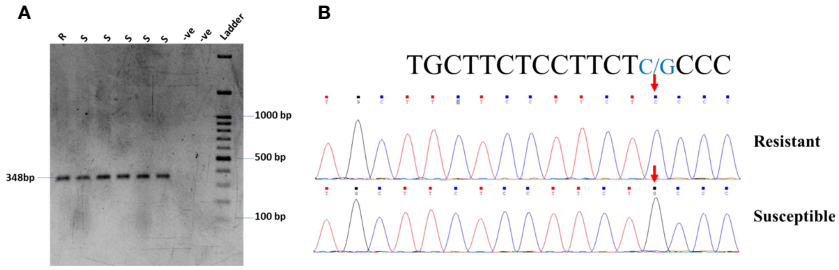


FIGURE 5
(A) PCR amplification of resistant and susceptible germplasm for gene *Glyma.18G25100* validation. **(B)** Chromatogram of *Glyma.18G25100* showing G/C transversion in soybean germplasm.

4 Discussion

Soybean is one of the most important crops grown in many countries (Rathore et al., 2021). However, Pakistan grows soybeans in a negligible area. Many reasons lead to the low cultivation and production of this crop in the country. Among these, YMD is the main problem in soybean production in the region, based on typical symptoms of YMD observed in the field in Pakistan. YMD also has an important role and impact on the production of mungbean in India as well as in Pakistan (Ilyas et al., 2009; Nair et al., 2017; Ashok, 2018). In India, this disease not only infects mungbean, but also poses a serious threat to other legume crops as well as to soybean production (Gazala et al., 2013). However, India has applied multiple approaches to find ways to control YMD in soybean crops, such as producing resistant cultivars (Mishra et al., 2020). However, despite the dire need for the cultivation of soybeans in Pakistan, there are no reports available regarding cultivars of soybeans resistant to YMD in the country.

Multiple approaches are applied for the control of begomoviruses that cause multiple diseases in plants. One such method is to control their vectors, such as whiteflies, which not only spread the viral disease but also suck plant saps and damage plant growth. Heavy insecticide spray is being used to control this pest at different plant growth stages. Many such insecticide sprays are not safe for humans and lead to pollution as they persist for a long time in the environment (Zikankuba et al., 2019). Another approach is to identify/develop resistant germplasm, which could be high-yielding and will be the source of resistance in future breeding programs (Snehi et al., 2015). The identification of these resistant cultivars is of high importance as these cultivars have less impact on the cost of production as compared to insecticide sprays, which are more expensive for the control of vectors transmitting these viruses (Habib et al., 2007). This approach was applied in mungbean to identify different resistant cultivars against YMD (Habib et al., 2007). The marker-assisted breeding approach was also used to screen and identify the YMD-resistant and high-yielding mungbean cultivars, which helped the farmers to grow mungbean (Binyamin et al., 2015). The success of conventional breeding for the control of YMD in mungbean encouraged us to acquire soybean germplasm from national and international sources and to identify virus-resistant germplasm. We screened the germplasm for a period of 5 years (2016-2020) under field conditions in the autumn season (August-November) when the incidence of yellow mosaic disease was at its maximum, and successfully identified YMD-resistant germplasm. This would be highly helpful for the soybean breeding program in the future.

The accessions were reduced with selected lines in successive generations by selecting only the HS, R, and HR soybean germplasm, but at the same time, new entries were also added to obtain the best possible gene pool for disease resistance. All the lines

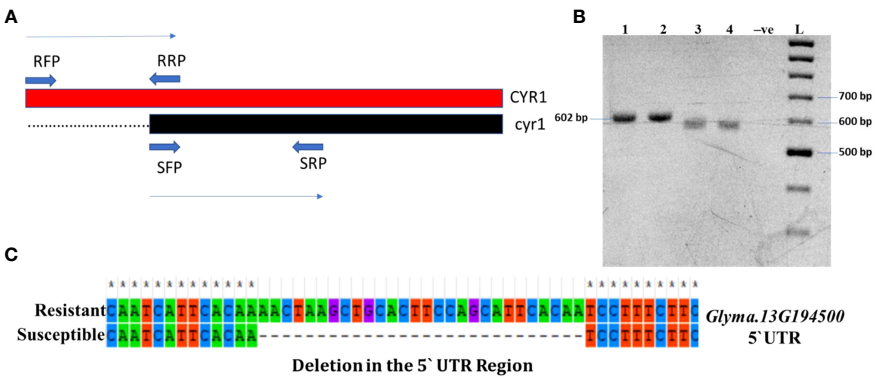


FIGURE 6
(A) Representation of overlapping primers designed to amplify complete *CYR-1* gene; *Glyma.13G194500* **(B)** PCR amplification of *CYR-1* gene; *Glyma.13G194500* **(C)** Alignment of *Glyma.13G194500* of YMD resistant and susceptible soybean. The symbol * shows homology.

were screened at least for three years. During the experiments, the populations of breeding and mutants of locally established cultivars were also developed (results not shown). These plant materials were also tested for YMD resistance (Table 3). Initially, the germplasm imported from the USDA along with the local cultivar (NARC-2021) were screened. It has been observed that single plant selection (SPS) is very important for disease resistance in crops (Foxe, 1992). So, for this reason, the SPS was also performed in the same cultivar having high disease severity, and all single plants resistant to YMD were selected and screened in the next generations.

Although symptoms on plant leaves are the initial indication of viruses, these symptoms are not reliable for the exact identification of viral strains or species in plants until the whole genome sequencing of the viral genome is performed (González-Garza, 2017) as the YMD is caused by four different species of LYMVs, namely, MYMIV, MYMV, HgYMV, and DoYMV (Qazi et al., 2007). So, molecular characterization of these species causing YMD in soybeans was needed before further evaluation. For this reason, the total DNA was extracted from both resistant and susceptible soybean samples and was characterized for MYMIV and MYMV. Both the viral species were identified in susceptible cultivars and there was no identification of other LYMVs (Table 2). This shows that both species (MYMIV and MYMV) are responsible for the YMD in soybean at the tested location (Faisalabad, Pakistan), and the data have been reported (Rahman et al., 2023a). The sequences for each of the species MYMIV and MYMV were submitted in NCBI with accession nos. MN885463 and MK098184, respectively.

Regarding YMD severity, the highest severity was found in many soybean cultivars such as NBG-22, NBG-31, NBG-47, NBG-117, NARC-2021, SPS-C1, and SPS-C9 (Table 3). In 2018, more germplasm were added which were developed by hybridization with local cultivars and mutation. These were given the code CF in cross and M for mutant seeds. The YMD incidence was higher in 2016 and 2019, and lower in 2018 and 2020 (Figure 2). The increase in disease incidence in 2019 was due to new soybean entries, whereas the decrease in disease incidence was due to recurrent selections of advanced disease-resistant soybean material throughout the period (2016-2020).

In the studied soybean lines, initially, no germplasm were HR for YMD and there was one single plant in one germplasm (cv. Jack) that was HR to YMD. However, after 2016, many lines showed complete immunity to YMD but the majority of these HR germplasm lost the resistance in the next generations, which shows that, initially, it was pseudo-resistance. Only one line cultivar named "SG-soybean", an SPS from cv. Jack in the year 2016, maintained complete resistance in respective generations (Table 3). This shows that the resistance in the SG-soybean line is stable. Stable resistance is very important in plant breeding programs (Stuthman et al., 2007), although in most cases the resistance is not stable and is lost due to the emergence of new viral strains/species, which is a routine phenomenon that viruses use for their survival and transmission. Hence, this change in the viral genome leads to the production of highly resistant viral strains/pathogens. This has been shown by the appearance of MYMIV in India, which is a novel species of LYMVs. Before the appearance of MYMIV, MYMV was the dominant species that caused YMD, and the resistance was lost due to the emergence of this novel species of MYMIV. However, it could be possible to identify the new resistance source against these two species of LYMVs. Similarly, the resistance break was also observed in cotton for CLCuD by the emergence of new strains of the cotton leaf curl virus, which caused epidemics in Pakistan (Zubair et al., 2017). In tomatoes, it has been observed that the reassortment of the tomato spotted wilt virus led to new strains and broke previously resistant cultivars (Margaria et al., 2015). In our previous investigation, there were many strains and species of begomoviruses identified, including both MYMIV and MYMV (Rahman et al., 2023a; Rahman et al., 2023b), however, the SG-soybean line was found resistant to high disease pressure and there were no symptoms on the SG-soybean line (2016-2020) in all generations (Figure 4), although the disease vector, namely, whiteflies, were present in a soybean field, which further strengthens the claim of stable resistance in the SG-soybean line. Second, the presence of MYMIV and MYMV in the field confirms the resistant cultivar having stable resistance. Another possibility of stable resistance could be the presence of the expression of resistance genes that interact and degrade viral proteins. The most resistant protein has been identified for MYMIV in black gram, which interacts with the rep protein of the virus (Patel et al., 2016). The resistance in plants is also governed by gene silencing. It has been found that the resistant cultivar induces viral RNA degradation earlier than the susceptible cultivars after infection (Yadav et al., 2009). Based on previous literature, it is predicted that the SG-soybean line also has some resistance genes that lead to resistance or the cultivar is resistant to whitefly. Further research is needed to determine the mechanism of resistance in the SG-soybean line.

YMD has a significant impact on plant yield and yield-linked agronomic traits (Baghel et al., 2010). Therefore, in the current study, the impact of YMD on soybeans was recorded based on the agronomic parameters (DF, DM, and PH) and yield-related parameters (PPP, SPP, and GY) of susceptible soybeans compared to resistant soybeans. The potential yield of NARC-2021 is 3000kg/ha. The yield of the SG-soybean line was found to be 7.9 times higher as compared to NARC-2021 and 3 times higher as compared to NBG-117 (Table 3). In Pakistan, the Nuclear Institute for Agriculture and Biology (NIAB) produced YMD-resistant cultivars of mungbean. High-yielding mungbean cultivars with susceptibility were crossed with resistant cultivars to further increase the yield (Khattak et al., 2008). This approach could be used to produce YMD-resistant and high-yielding soybean cultivars. The SG-soybean is of short stature (37.3 ± 2.0a cm), having purple flowers, brown pods, golden yellow seeds, off-black hilum, and narrow leaves with dark green color (Figure 7). The dark color is a clear indication that the phytochemicals are high, which is of high importance in disease resistance. Further studies are needed to investigate the phytochemical profiles of YMD-resistant and susceptible soybean cultivars.

After the identification of resistant and susceptible cultivars through screening, molecular identification of host genes is very

important for a breeding program. There are no such reports in soybeans for YMD resistance genes (Singh et al., 1974). Yadav et al. (2015) identified a SNP that leads to resistance against MYMIV. They found that the mutation, which is a transversion of C to G (*Glyma.18G25100)*, leads to the susceptibility of soybeans to YMD. We hypothesized that the same mutation is responsible for disease susceptibility in tested germplasm. To investigate this, the highly resistant and susceptible genotypes of soybeans were tested for the said mutation. We identified the same SNP in the *Glyma.18G025100* gene on chromosome 18 in the susceptible cultivars; however, this mutation was missing in some susceptible cultivars (Figure 5B). It has also been reported that monogenic resistance is of high importance in the initial stages of infection, but in most cases, the monogenic resistance is not durable as the pathogens mutate DNA for their survival (Stuthman et al., 2007).

In 2016, Patel et al. (2016) reported a deletion of 128 amino acids at the start of the dominant *CYR-1* allele that leads to protein truncation and susceptibility to YMD. We, therefore, hypothesized that the ortholog of the *CYR-1* gene is present in soybeans, which may also lead to YMD susceptibility. To test this hypothesis, the sequence of the *CYR-1* gene was aligned with the whole-genome sequence (WGS) of soybeans. We found the closest match to be the 65% sequence similarity of the soybean gene, *Glyma.13G194500*, which was selected for further study. We used overlapping primers to amplify the ortholog in the resistant and susceptible soybean germplasm. Interestingly, we identified a novel deletion of 28 bp in the 5`UTR of the *Glyma.13G194500* gene in only one susceptible soybean accession (Figure 6C), while this deletion was not observed in other susceptible cultivars. We named the gene *Glyma.13G194500* as *cyr-1*. This deletion was something unusual, and based on the observation it was expected that the RNA would also be truncated, which would lead to truncated protein or no expression (no protein synthesis). To test this, the total RNA was extracted to synthesize the cDNA, which was used as a template to amplify the *cyr-1* gene transcripts. There was no amplification in the susceptible cultivar, whereas transcripts were amplified in the resistant cultivar. This showed that the RNA is not synthesized in the susceptible cultivar, which may lead to susceptibility in soybeans. Western blotting is also needed to further confirm the absence of the resulting protein.

The bioinformatics analysis by PlantPAN3.0 (Chow et al., 2019) revealed the presence of 14 TFBS spots common in both 5`UTR regions from the CYR-1 gene from resistant and susceptible germplasm, whereas one binding site, bHLH, was additionally found in the 5`UTR region of the CYR-1 gene of resistant germplasm but was absent in the 5`UTR of the CYR-1 gene in susceptible germplasm (Table S-4), which reveals the importance of the 5`UTR region in YMD resistance. The ProtParam tool detected slight differences in some parameters of both proteins (Gasteiger et al., 2005). Both proteins have 644 amino acid length, however, resistant CYR-1 has 70805.07 Da molecular weight while susceptible CYR-1 protein has 70779.03 Da. Other features have been highlighted in Table S-5. Protein structures obtained from I-TASSER (Zhou et al., 2022) and TM-align (Zhang and Skolnick, 2005) visualized in PyMOL are shown in Figures 8A–D. Both proteins from resistant and susceptible germplasm were superimposed and aligned, which showed a 0.87219 TM-score and root-mean-square deviation (RMSD) score of 3.72, displaying the same folds for both proteins. These results conveyed that the proteins with SNP differences only showed minor variations in structure and parameters. Furthermore, the in-silico interaction of CYR-1 both from resistant and susceptible germplasm with the MYMIV viral coat protein exhibited a 0.3537 PSOPIA score (Murakami and Mizuguchi, 2014). Binding affinity in terms of ΔΔG values was detected at the same rate of -10.861 for both proteins through ISLAND (Abbasi et al., 2020), showing no difference in binding affinity of both resistant and susceptible CYR-1 protein to the viral coat protein, which could further be subjected to validation in future studies.

The present investigation has prime importance to uplift soybean research and cultivation in the country. The resistant germplasm could be used to transfer resistance to susceptible cultivars. Identified susceptible and resistant cultivars could also be used as a check in YMD screening experiments. Faisalabad is a hub for YMD incidence under natural field conditions, so the

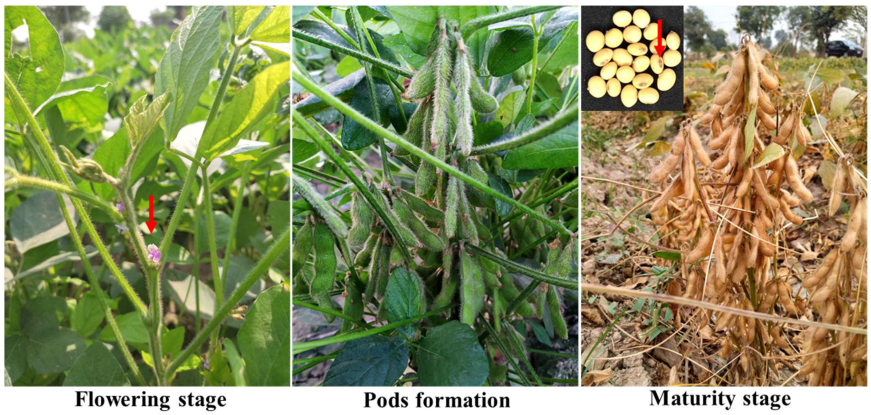


FIGURE 7
Different growth stages of resistant germplasm, NBG-SG-soybean, in the field.

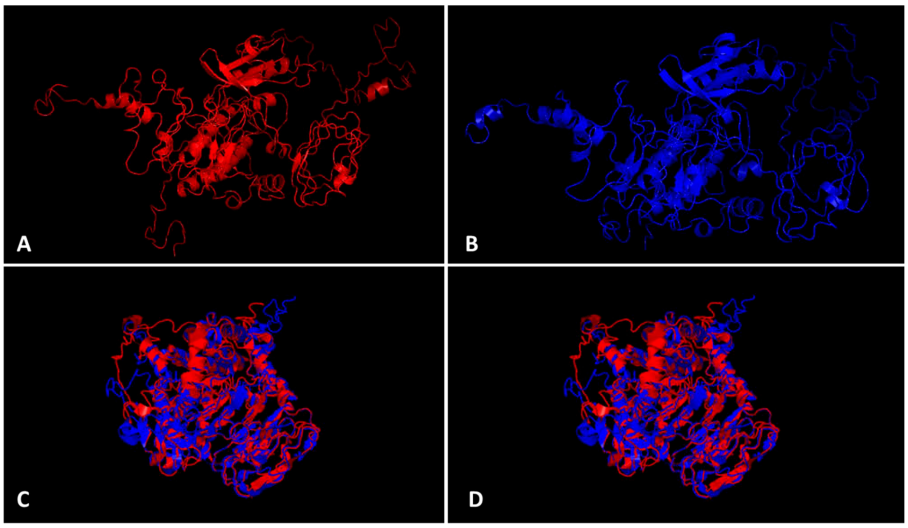


FIGURE 8
CYR-1 protein structure prediction and alignment. Protein structures identified through I-TASSER visualized in PyMOL **(A)** CYR-1 from resistant and **(B)** CYR-1 from susceptible germplasm, **(C, D)** superimposed and aligned *CYR-1* proteins, which showed a 0.87219 TM-score and root-mean-square deviation (RMSD) score of 3.72, displaying same folds for both proteins.

resistant cultivars at this location will be of high importance (Sudha et al., 2013). Moreover, the SNP and *CYR-1* ortholog could be used in marker-assisted breeding to screen YMD-resistant and susceptible soybean germplasm. In the future, the YMD chart (Figure 1) developed in the present study can be used for disease scoring in soybeans, facilitating the work of plant virologists to measure disease severity with accuracy.

To the best of our knowledge, this is the first report on the identification of resistant/susceptible cultivars and molecular marker identification against YMD in soybean cultivars.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

SM, ZM and GR gave the idea, supervised the experiments, and review the final draft of the manuscript. SR, GR, EM and RZN performed the experiments, analyzed the data, and wrote the first draft. The protein *In-silico* study was performed by RZN. IA helped in data analysis. WAP and PL helped in Sanger sequencing of resistant and susceptible soybean germplasm and bioinformatics analysis. MH helped in field data collection. AZG and MSH performed the experiments, analyzed the data, and wrote the first draft of manuscript. Also helped in Sanger sequencing of resistant and susceptible soybean germplasm and bioinformatics analysis.

Funding

The present research work was supported by the Punjab Agriculture Research Board (PARB) under Project# 830. The sequencing of whole genes by chromosome walking and finding SNP in soybean was supported by HEC-Pakistan under IRSIP at the University of Georgia, Athens, USA.

Acknowledgments

We are thankful to the National Agriculture Research Center (NARC), Islamabad, Pakistan, Ayub Agriculture Research Institute (AARI), and USDA, USA for providing soybean genotypes. The authors extend their appreciation to the Researchers Supporting Project number RSPD2023R686, King Saud University, Riyadh, Saudi Arabia.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1230559/full#supplementary-material

SUPPLEMENTARY TABLE 1
Field screening of soybean germplasm for YMD resistance in 2016-2020.

SUPPLEMENTARY TABLE 2
Primers for YMD screening in the soybean field.

SUPPLEMENTARY TABLE 3
Primers for the amplification of *CYR-1* and *cyr-1* gene ortholog in soybean.

SUPPLEMENTARY TABLE 4
TFBSs identified in 5`UTR of *CYR-1* gene in Resistant and susceptible germplasm.

SUPPLEMENTARY TABLE 5
CYR1 proteins' physio-chemical features.

References

Abbasi, W. A., Yaseen, A., Hassan, F. U., Andleeb, S., and Minhas, F. U. A. A. (2020). ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Min.* 13 (1), 1–13. doi: 10.1186/s13040-020-00231-w

Ashok, K. (2018). Management of yellow mosaic disease of soybean. *Int. J. Agric. Sci.* 10 (16), 6913–6915.

Baghel, G., Jahan, T., Afreen, B., Naqvi, Q., Snehi, S., and Raj, S. (2010). Detection of a begomovirus associated with yellow mosaic disease of Ashwagandha (*Withania somnifera*) and its impact on biomass yield. *Med. Plants Int. J. Phytomed.* 2, 219–223. doi: 10.5958/j.0975-4261.2.3.034

Basak, J., Kundagrami, S., Ghose, T., and Pal, A. (2005). Development of yellow mosaic virus (YMMV) resistance linked DNA marker in Vigna mungo from populations segregating for YMMV-reaction. *Mol. Breed.* 14, 375–383. doi: 10.1007/s11032-005-0238-6

Binyamin, R., Khan, M., Khan, N., and Khan, A. (2015). Application of SCAR markers linked with mungbean yellow mosaic virus disease-resistance gene in Pakistan mungbean germplasm. *Genet. Mol. Res.* 14, 2825–2830. doi: 10.4238/2015.March.31.13

Chow, C. N., Tzong, Y. L., Yu-Cheng, H., Guan-Zhen, L., Kuan-Chieh, T., Ya-Hsin, L., et al. (2019). PlantPAN3.0: a new and updated resource for reconstructing transcriptional regulatory networks from ChIP-seq experiments in plants. *Nucleic Acids Res.* 47 (D1), D1155–D1163. doi: 10.1093/nar/gky1081

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19, 11–15.

Foxe, M. (1992). Breeding for viral resistance: conventional methods. *Neth. J. Plant Pathol.* 98, 13–20. doi: 10.1007/BF01974467

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. E., Wilkins, M. R., Appel, R. D., et al. (2005). *Protein identification and analysis tools on the ExPASy server* (Totowa, New Jersey, United States: Humana press), 571–607.

Gazala, I. S., Sahoo, R., Pandey, R., Mandal, B., Gupta, V., Singh, R., et al. (2013). Spectral reflectance pattern in soybean for assessing yellow mosaic disease. *Indian. J. Virol.* 24, 242–249. doi: 10.1007/s13337-013-0161-0

González-Garza, R. (2017). Evolution of diagnostic technics for plant viruses. *Rev. Mex. Fitopatol.* 35, 591–610. doi: 10.18781/r.mex.fit.1706-1

Gupta, S., Gupta, D. S., Anjum, T. K., Pratap, A., and Kumar, J. (2013). Inheritance and molecular tagging of MYMIV resistance gene in blackgram (Vigna mungo L. *Hepper) Euphytica.* 193, 27–37. doi: 10.1007/s10681-013-0884-4

Habib, S., Shad, N., Javaid, A., and Iqbal, U. (2007). Screening of mungbean germplasm for resistance/tolerance against yellow mosaic disease. *Mycopath* 5 (2), 89–94.

Ilyas, M., Qazi, J., Mansoor, S., and Briddon, R. W. (2009). Molecular characterisation and infectivity of a "Legumovirus" (genus Begomovirus: family Geminiviridae) infecting the leguminous weed *Rhynchosia minima* in Pakistan. *Virus. Res.* 145, 279–284. doi: 10.1016/j.virusres.2009.07.018

Ilyas, M., Qazi, J., Mansoor, S., and Briddon, R. W. (2010). Genetic diversity and phylogeography of begomoviruses infecting legumes in Pakistan. *J. Gen. Virol.* 91, 2091–2101. doi: 10.1099/vir.0.020404-0

Islam, S., Munshi, A., Mandal, B., Kumar, R., and Behera, T. (2010). Genetics of resistance in *Luffa cylindrica Roem.* against Tomato leaf curl New Delhi virus. *Euphytica* 174, 83–89. doi: 10.1007/s10681-010-0138-7

Khattak, G. S. S., Saeed, I., and Shah, S. A. (2008). Breeding high yielding and disease resistant mungbean (Vigna radiata (L.) Wilczek) genotypes. *Pak. J. Bot.* 40, 1411–1417.

Kitsanachandee, R., Somta, P., Chatchawankanphanich, O., Akhtar, K. P., Shah, T. M., Nair, R. M., et al. (2013). Detection of quantitative trait loci for mungbean yellow mosaic India virus (MYMIV) resistance in mungbean (Vigna radiata (L.) Wilczek) in India and Pakistan. *Breed. Sci.* 63, 367–373. doi: 10.1270/jsbbs.63.367

Lal, S., Rana, V., Sapra, R., and Singh, K. (2005). Screening and utilization of soybean germplasm for breeding resistance against Mungbean Yellow Mosaic Virus. *Soybean Genet. News Lett.* 1, 32.

Maiti, S., Basak, J., Kundagrami, S., Kundu, A., and Pal, A. (2011). Molecular marker-assisted genotyping of mungbean yellow mosaic India virus resistant germplasms of mungbean and urdbean. *Mol. Biotechnol.* 47, 95–104. doi: 10.1007/s12033-010-9314-1

Maiti, S., Paul, S., and Pal, A. (2012). Isolation, characterization, and structure analysis of a non-TIR-NBS-LRR encoding candidate gene from MYMIV-resistant *Vigna mungo*. *Mol. Biotechnol.* 52, 217–233. doi: 10.1007/s12033-011-9488-1

Margaria, P., Ciuffo, M., Rosa, C., and Turina, M. (2015). Evidence of a tomato spotted wilt virus resistance-breaking strain originated through natural reassortment between two evolutionary-distinct isolates. *Virus Res.* 196, 157–161. doi: 10.1016/j.virusres.2014.11.012

Maruthi, M., Manjunatha, B., Rekha, A., Govindappa, M., Colvin, J., and Muniyappa, V. (2006). Dolichos yellow mosaic virus belongs to a distinct lineage of Old World begomoviruses; its biological and molecular properties. *Ann. Appl. Biol.* 149, 187–195. doi: 10.1111/j.1744-7348.2006.00075.x

Mishra, G. P., Dikshit, H. K., Tripathi, K., Kumar, R. R., Aski, M., Singh, A., et al. (2020). Yellow mosaic disease (YMD) of mungbean (Vigna radiata (L.) Wilczek): current status and management opportunities. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00918

Murakami, Y., and Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using averaged one-dependence estimators. *BMC Bioinf.* 15 (1), 1–11. doi: 10.1186/1471-2105-15-213

Nair, R., Gotz, M., Winter, S., Giri, R., Boddepalli, V., Sirari, A., et al. (2017). Identification of mungbean lines with tolerance or resistance to yellow mosaic in fields in India where different begomovirus species and different Bemisia tabaci cryptic species predominate. *Eur. J. Plant Pathol.* 149, 349–365. doi: 10.1007/s10658-017-1187-8

Patel, A., Maiti, S., Kumar, S., Ganguli, S., and Pal, A. (2016). An integrated approach to comprehend MYMIV-susceptibility of blackgram Cv. T9 possessing allele of CYR1, the cognate R-gene. *Am. J. Plant Sci.* 7, 267. doi: 10.4236/ajps.2016.72026

Qazi, J., Ilyas, M., Mansoor, S., and Briddon, R. W. (2007). Legume yellow mosaic viruses: genetically isolated begomoviruses. *Mol. Plant Pathol.* 8, 343–348. doi: 10.1111/j.1364-3703.2007.00402.x

Rahman, S. U., Domier, L. L., Raza, G., Ahmed, N., McCoppin, N. K., Amin, I., et al. (2023b). Metagenomic study for the identification of viruses infecting soybean in Pakistan. *Australas. Plant Pathol.* 52, 191–194. doi: 10.1007/s13313-023-00909-9

Rahman, S. U., Raza, G., Zubair, M., Ahmed, N., Domier, L. L., Jamil, N., et al. (2023a). Multiple begomoviruses infecting soybean; a case study in Faisalabad, Pakistan. *Biologia* 78 (2), 609–620. doi: 10.1007/s11756-022-01290-6

Ram, H. H., Singh, K., and Verma, V. (1984). Breeding for resistance to yellow mosaic virus through interspecific hybridization in soybean. *Soybean Genet. Newslett.* 11, 46–48.

Rani, A., Kumar, V., Rathi, P., and Shukla, S. (2017). Linkage mapping of Mungbean yellow mosaic India virus (MYMIV) resistance gene in soybean. *Breed. Sci.* 67, 95–100. doi: 10.1270/jsbbs.16115

Rathore, V., Sharma, H., and Narvariya, R. (2021). Growth rate of cost of cultivation of soybean in Maharashtra States of India. *Pharm. Innov. J.* 10 (3), 84–89. doi: 10.22271/tpi.2021.v10.i3Sb.5838

Singh, B., Singh, B., and Gupta, S. (1974). PI 171.443 and G. formosana-resistant lines for yellow mosaic of soybean. *Soybean Genet. Newslett.* 1, 17–18.

Snehi, S., Raj, S., Prasad, V., and Singh, V. (2015). Recent research findings related to management strategies of begomoviruses. *J. Plant Pathol. Microbiol.* 6, 6. doi: 10.4172/2157-7471.1000273

Souframanien, J., and Gopalakrishna, T. (2006). ISSR and SCAR markers linked to the mungbean yellow mosaic virus (MYMV) resistance gene in blackgram [Vigna mungo (L.) Hepper]. *Plant Breed.* 125, 619–622. doi: 10.1111/j.1439-0523.2006.01260.x

Stuthman, D., Leonard, K., and Miller-Garvin, J. (2007). Breeding crops for durable resistance to disease. *Adv. Agron.* 95, 319–367. doi: 10.1016/S0065-2113(07)95004-X

Sudha, M., Karthikeyan, A., Nagarajan, P., Raveendran, M., Senthil, N., Pandiyan, M., et al. (2013). Screening of mungbean (Vigna radiata) germplasm for resistance to Mungbean yellow mosaic virus using agroinoculation. *Can. J. Plant Pathol.* 35, 424–430. doi: 10.1080/07060661.2013.827134

Varma, A., and Malathi, V. (2003). Emerging geminivirus problems: a serious threat to crop production. *Ann. Appl. Biol.* 142, 145–164. doi: 10.1111/j.1744-7348.2003.tb00240.x

Wrather, J. A., Anderson, T., Arsyad, D., Gai, J., Ploper, L., Porta-Puglia, A., et al. (1997). Soybean disease loss estimates for the top 10 soybean producing countries in 1994. *Plant Dis.* 81, 107–110. doi: 10.1094/PDIS.1997.81.1.107

Yadav, C. B., Bhareti, P., Muthamilarasan, M., Mukherjee, M., Khan, Y., Rathi, P., et al. (2015). Genome-wide SNP identification and characterization in two soybean cultivars with contrasting mungbean yellow mosaic India virus disease resistance traits. *PloS One* 10, e0123897. doi: 10.1371/journal.pone.0123897

Yadav, R. K., Shukla, R. K., and Chattopadhyay, D. (2009). Soybean cultivar resistant to Mungbean Yellow Mosaic India Virus infection induces viral RNA degradation earlier than the susceptible cultivar. *Virus. Res.* 144, 89–95. doi: 10.1016/j.virusres.2009.04.011

Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 33 (7), 2302–2309. doi: 10.1093/nar/gki524

Zhou, X., Zheng, W., Li, Y., Pearce, R., Zhang, C., Bell, E. W., et al. (2022). I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nat. Protoc.* 17, 2326–2353. doi: 10.1038/s41596-022-00728-0

Zikankuba, V. L., Mwanyika, G., Ntwenya, J. E., and James, A. (2019). Pesticide regulations and their malpractice implications on food and environment safety. *Cogent. Food. Agric.* 5, 1601544. doi: 10.1080/23311932.2019.1601544

Zubair, M., Zaidi, S.-e., Shakir, S., Farooq, M., Amin, I., Scheffler, J. A., et al. (2017). Multiple begomoviruses found associated with cotton leaf curl disease in Pakistan in early 1990 are back in cultivated cotton. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-00727-2

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 December 2023
DOI 10.3389/fpls.2023.1308731

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Guizhen Kan,
Nanjing Agricultural University, China
Yuri Shavrukov,
Flinders University, Australia

*CORRESPONDENCE
Zenglu Li
zli@uga.edu

RECEIVED 07 October 2023
ACCEPTED 28 November 2023
PUBLISHED 20 December 2023

CITATION
Souza R, Rouf Mian MA, Vaughn JN and Li Z (2023) Introgression of a Danbaekkong high-protein allele across different genetic backgrounds in soybean. *Front. Plant Sci.* 14:1308731. doi: 10.3389/fpls.2023.1308731

Introgression of a Danbaekkong high-protein allele across different genetic backgrounds in soybean

Renan Souza[1], M. A. Rouf Mian[2], Justin N. Vaughn[1,3] and Zenglu Li[1]*

[1]Department of Crop and Soil Sciences, University of Georgia, Athens, GA, United States, [2]Soybean and Nitrogen Fixation Research Unit, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), Raleigh, NC, United States, [3]Genomics and Bioinformatics Research Unit, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), Athens, GA, United States

Soybean meal is a major component of livestock feed due to its high content and quality of protein. Understanding the genetic control of protein is essential to develop new cultivars with improved meal protein. Previously, a genomic region on chromosome 20 significantly associated with elevated protein content was identified in the cultivar Danbaekkong. The present research aimed to introgress the Danbaekkong high-protein allele into elite lines with different genetic backgrounds by developing and deploying robust DNA markers. A multiparent population consisting of 10 F_5-derived populations with a total of 1,115 recombinant inbred lines (RILs) was developed using "Benning HP" as the donor parent of the Danbaekkong high-protein allele. A new functional marker targeting the 321-bp insertion in the gene *Glyma.20g085100* was developed and used to track the Danbaekkong high-protein allele across the different populations and enable assessment of its effect and stability. Across all populations, the high-protein allele consistently increased the content, with an increase of 3.3% in seed protein. A total of 103 RILs were selected from the multiparent population for yield testing in five environments to assess the impact of the high-protein allele on yield and to enable the selection of new breeding lines with high protein and high yield. The results indicated that the high-protein allele impacts yield negatively in general; however, it is possible to select high-yielding lines with high protein content. An analysis of inheritance of the Chr 20 high-protein allele in Danbaekkong indicated that it originated from a *Glycine soja* line (PI 163453) and is the same as other *G. soja* lines studied. A survey of the distribution of the allele across 79 *G. soja* accessions and 35 *Glycine max* ancestors of North American soybean cultivars showed that the high-protein allele is present in all *G. soja* lines evaluated but not in any of the 35 North American soybean ancestors. These results demonstrate that *G. soja* accessions are a valuable source of favorable alleles for improvement of protein composition.

KEYWORDS

soybean, seed protein, Danbaekkong, chromosome 20 QTL, multiparent populations, yield

1 Introduction

Soybean [*Glycine max* (L.) Merr.] is one of the most important sources of protein and oil for direct and indirect human use. Soybean oil is omnipresent in the food industry, while soybean meal is the primary source of protein for livestock. Over the past 33 years, soybean yield in the United States increased 40.8%; however, the protein content went in an opposite direction, decreasing from 35.8% to 33.5% (Naeve and Miller-Garvin, 2021). Farmers and grain processors historically have had no incentive to produce and deliver soybeans with high protein and therefore no focus has been given in improving this seed component. The reduction in seed protein has negative effects on soybean value, as lower protein content makes it difficult to meet the requirements of the livestock industry for feed (Brumm and Hurburgh, 1990; de Borja Reis et al., 2020).

The genetic component is a major factor in the determination of seed composition in soybean. Lee et al. (2019) demonstrated the importance of the genetic factors for protein composition (heritability of 0.94) and confirmed the antagonist relationship between protein and oil ($r = -0.75$; $p < 0.0001$). More than 160 protein quantitative trait loci (QTLs) from 35 different studies have been reported (Patil et al., 2017) and one of these QTLs, located on chromosome (Chr) 20, has been repeatedly identified in several studies (Diers et al., 1992; Hwang et al., 2014; Vaughn et al., 2014; Warrington et al., 2015; Qi et al., 2016). This QTL has received the attention of many researchers because of its high additive effect and stability (Lestari et al., 2013). Warrington et al. (2015) demonstrated that this QTL explained 55% of the phenotypic variation of seed protein content in a bi-parental population derived from a cross of "Benning" (PI 595645) (Boerma et al., 1997) and Danbaekkong (PI 619083) (Kim et al., 1996). Danbaekkong is a South Korean cultivar that contributed to high protein content in the population (Warrington et al., 2015).

Despite the negative relationship of protein with oil and yield, there were reports on the feasibility of developing lines with increased protein content and high yield (Cober and Voldeng, 2000; Brzostowski et al., 2017). Prenger et al. (2019) developed Benning HP as a near-isogenic line (NIL) with a high-protein allele on Chr 20 by backcrossing an F_5-derived line from Benning × Danbaekkong to the recurrent parent Benning. This line has high protein content and yield equivalent to the recurrent parent Benning, demonstrating that it is possible to mitigate the negative effects of the high-protein allele on yield with progeny selection. However, it is still not clear how the protein and yield relationship work in multiple genetic backgrounds.

A Chr 20 QTL for protein content was detected in the same location of previous mapping studies in a genome-wide association study (GWAS) with accessions from the USDA Soybean Germplasm Collection conducted by Vaughn et al. (2014). Bandillo et al. (2015) also analyzed 12,000 accessions from the same collection and identified a protein QTL in the same region. The GWAS hits in these studies were associated with the alleles frequently found in Korean accessions. Using the similar dataset, Patil et al. (2017) performed a genome-wide phylogenetic analysis comparing Danbaekkong, North American Soybean Ancestors (NASA), Asian landraces, and several *Glycine soja* lines. When all SoySNP50K SNPs were considered, Danbaekkong was clustered with the NASA; however, when SNPs in the range of 27–32 Mb on Chr 20 were analyzed, Danbaekkong was clustered separately from NASA. This result indicated that NASA likely have a different allele from Danbaekkong at the Chr 20 and introgression of the high-protein allele into elite soybean lines could improve the seed protein content.

Soybean accessions in the USDA Germplasm Collection have great variation for protein content, with accessions reaching up to 57% of seed protein (USDA, 2023). This resource can be tapped to increase the overall protein content and quality in soybean breeding programs. It has been observed that *G. max* cultivars developed in Asia, especially in South Korea, usually have a higher content of seed protein than those developed in other countries (Vaughn et al., 2014; Bandillo et al., 2015; Patil et al., 2017). It is likely a result of the historical breeding efforts in that region to focus on the improvement of seed composition for soy food products, such as tofu and soy sauce (Lee et al., 2015). Danbaekkong is a cultivar developed in South Korea based on the selection for seed yield, protein content, quality, and tofu yield (Kim et al., 1996). The Korean accessions with high protein content are an important source of genetic diversity that can be used in U.S. soybean breeding programs to improve nutritional composition.

Recently, a gene was identified underlying control of the protein QTL on Chr 20. Fliege et al. (2022) performed fine mapping in multiple populations using a *G. soja* line (PI 468916) as the QTL donor and narrowed the QTL interval to a region of 77.8 kb. In this interval, a 321-bp fragment was present in the 4th exon of the gene G*lyma.20g085100* in low-protein lines. Using an RNAi experiment, the authors demonstrated that the variation in *Glyma.20g085100* was responsible for the difference in protein content. Similarly, Goettel et al. (2022) indicated that *Glyma.20g085100* is the gene responsible for elevated protein at the Chr 20 QTL and soybean lines without the 321-bp insertion exhibit increased protein content, while the lines with the 321-bp insertion had low protein. The authors concluded that the insertion was likely caused by a transposable element, and during the domestication process, the insertion allele is fixed in most *G. max* lines.

In the present research, we aimed to validate the Chr 20 QTL from Danbaekkong for increased protein content, introgress the allele into a wide range of genetic backgrounds for protein improvement, and elucidate the inheritance of the Danbaekkong high-protein allele.

2 Materials and methods

2.1 Plant materials and population development

The population consisting of 140 recombinant inbred lines (RILs) derived from Benning × Danbaekkong originally used to map the Chr 20 QTL was analyzed to saturate the QTL region. The seeds, original phenotypic data, and genotypic data were obtained from Warrington et al. (2015). To enable identification of polymorphisms in the QTL

region, seven soybean lines with high and low protein content were selected for genome sequencing (Supplementary Table S1). The elite parent Benning and the high-protein parent Danbaekkong (PI 619083) were sequenced together with one high-protein *G. soja* accession (PI 163453) and three high-protein *G. max* lines (PI 398589, PI 408012, and PI 602447) that have a haplotype in the QTL region similar to Danbaekkong. The sequence of the *G. soja* accession PI 468916 that was used in the original mapping study of the Chr 20 QTL by Diers et al. (1992) was obtained from Zhou et al. (2015) and Bayer et al. (2021).

A set of 10 populations was developed by crossing Benning HP with 10 elite lines in 2016 (Supplementary Table S2). Benning HP is a MG VII NIL of Benning (PI 595645), carrying the introgression of the Chr 20 high-protein allele from Danbaekkong (PI 619083) (Prenger et al., 2019). The populations have a structure of a nested association mapping population, where Benning HP is the hub parent (Supplementary Figure S1).

The F_1 generation was grown in the University of Georgia (UGA) greenhouse in Athens, GA during the winter of 2016–2017. During the summer of 2017, the F_2 generation was grown at the UGA Iron Horse Farm in Watkinsville, GA, and then two cycles of single seed descent advancement were conducted to advance the F_3 and F_4 generations during the winter of 2017–2018 in the Puerto Rican nursery. In 2018, the F_5 generation was grown at the UGA Iron Horse Farm and plants from each population were harvested and threshed individually. In the summer of 2019, plant rows were grown in an unreplicated augmented design along with the parents and three commercial check cultivars AG5534, AG6534, and AG7934.

2.2 Whole genome re-sequencing

The lines selected for sequencing were grown in a greenhouse and leaf tissue was collected 3 weeks after planting. For each genotype, a bulked sample of 12 plants were collected and leaf tissue was lyophilized and ground. Genomic DNA was extracted using the GeneJet Plant Genomic DNA purification mini kit (Thermo Scientific, Boston, MA, USA) and 150-bp DNA fragments were sequenced with the NextSeq Sequencing instrument (Illumina, San Diego, CA). Adapters were removed from the raw Fastq files using Trimmomatic v0.36 (Bolger et al., 2014), and sequencing reads were mapped to the soybean genome Wm82.a2.v1 (https://data.jgi.doe.gov) with Bowtie2 v2.3.3.1 (Langmead and Salzberg, 2012). SNP and indel calls were performed with the GATK HaplotypeCaller software (McKenna et al., 2010) and variants were annotated with SNPeff version 4.3t (Cingolani et al., 2012). Variant visualization in the Chr 20 QTL region was performed with the Integrative Genomics Viewer (IGV - v2.9.5) (Robinson et al., 2011).

2.3 Marker design and genotyping

The RILs from the Benning × Danbaekkong population were planted in the greenhouse and DNA extraction was performed on leaf tissue using the CTAB method (Keim et al., 1988). For the multiparent population, DNA was extracted from seed samples from all 1115 RILs in the 10 populations with a modified Edwards extraction (Edwards et al., 1991). KASP (LGC, Hoddesdon, UK) and TaqMan assays (Applied Biosystems, Foster City, CA) were designed using Geneious Primer version 2021.2 based on polymorphisms present in the QTL region identified from the SoySNP50K data (Song et al., 2013) and whole genome sequence of the seven sequenced soybean lines (Danbaekkong, Benning, PI 163453, PI 398589, PI 408012, PI 602447, and PI 468916) (Supplementary Tables S3–S5). The gene-specific marker GSM1252 targeting the 321-bp insertion at the gene *Glyma.20g085100* was designed based on information previously published by Fliege et al. (2022) and Goettel et al. (2022).

KASP reactions were performed in a 4-µL volume with 2 µL of master mix (1.97 µl of KASP 2X and 0.053 µL of primers) and 2 µL of 10–20 ng/µL genomic DNA. Similarly, TaqMan reactions were also conducted in a 4-µL volume including 2 µL of master mix (2 µL of TaqMan Universal Master Mix II and 0.2 µL of 5X Custom TaqMan SNP Genotyping Assay) and 2 µL of 10–20 ng/µL genomic DNA. PCR was performed in the BioRad C1000 Touch Thermal Cycler and PCR plates were read in either LightCycler® 480 (Roche, Germany) or TECAN infinite M200 microplate reader (Tecan US, Inc, Durham, NC) using the software KlusterCaller (version 2.24.0.11, LGC Genomics). Cycling conditions for the KASP assays were 15 min at 94°C, 10 cycles of 15 s at 94°C and 1 min at 65°C and 30 cycles of 20 s at 94°C and 1 min at 57°C. Cycling conditions for the TaqMan followed a modified touchdown PCR with an initial 10 min at 95°C, 10 cycles of 20 s at 95°C and 1 min at 71°C, decreasing 0.5°C each cycle, and 30 cycles of 15 s at 92°C and 1 min at 58°C.

2.4 Diversity panel

To analyze the distribution of the Chr 20 high-protein QTLs, 35 NASA (Gizlice et al., 1994) and 79 diverse *G. soja* accessions (La et al., 2019) were genotyped using the gene-specific TaqMan marker GSM1252. The 35 *G. max* soybean ancestors contributed 95% of the genes found in modern soybean cultivars (Gizlice et al., 1994) and the 79 *G. soja* lines are a core set that represent the genetic diversity within the entire USDA *G. soja* Collection (La et al., 2019). These accessions were planted in the greenhouse and leaf tissue was collected 2 weeks after planting. DNA extraction was performed with the CTAB method (Keim et al., 1988).

The seed composition of the 79 *G. soja* accessions was obtained from La et al. (2019) and the data for 25 of 35 North American Soybean ancestors were collected with the Near-Infrared Spectroscopy Perten DA 7250 Analyzer (PerkinElmer Inc., Waltham, MA, USA) from seeds harvested in the USDA winter nursery in Puerto Rico in 2018. The phenotypes of the remaining 10 accessions were retrieved from USDA GRIN (https://npgsweb.ars-grin.gov/gringlobal/).

Another panel of 35 *G. soja* lines was assembled to compare the genome sequence variation at the gene level and survey the distribution of the high-protein allele. The raw sequencing data were generated in previous studies (Bayer et al., 2021; Valliyodan et al., 2021) and are available at the Short Read Archive (SRA)

database at NCBI (www.ncbi.nlm.nih.gov). Adapters were removed from the raw Fastq files using Trimmomatic v0.36 (Bolger et al., 2014) and sequencing reads were mapped to the soybean genome Wm82.a2.v1 with Bowtie2 v2.3.3.1 (Langmead and Salzberg, 2012).

2.5 Yield trials of selected RILs

To understand the effects of the Danbaekkong high-protein allele on yield in different genetic backgrounds, a set of RILs from the multiparent populations with high and normal protein content were selected for evaluation in yield trials. A total of 103 lines were planted in three locations in 2020 and 2021. In 2020, all the lines were grown in a randomized complete block design with two replications per location and each line was planted in a 2-row plot with a length of 4.9 m spaced by 76.2 cm and a planting density of 27 seeds m^{-1}. A total of 46 lines were selected based on yield and agronomic performance and grown in 2021 in a randomized complete block design with three replications in a 4-row plot with the same plot length and row spacing. The commercial cultivars AG 64X8RR2X, AG 74X8RR2X, and AGS 738RR were used as checks across the different environments. Agronomic practices followed the recommended guidance for soybean production in Georgia (Bryant, 2020). All plots were end-trimmed before harvest to avoid edge effect, resulting in a length of 3.7 m. The two center rows were harvested, and weight and moisture were measured on combines. Approximately 200 seeds were sampled from each plot for seed composition analysis.

2.6 Seed composition analysis

The contents of protein and oil were determined using the NIR Perten DA 7250 Analyzer (PerkinElmer Inc., Waltham, MA, USA) and the instrument was calibrated by the manufacturer using thousands of samples with known seed composition values for whole seed and ground seed samples. Seed composition was reported on a dry matter basis. Analysis of the multiparent population was performed on the seeds from single plants in 2018 and from the plant rows in 2019. For the yield trials in 2020 and 2021, samples of 200 seeds were obtained from each plot.

2.7 Statistical and QTL analyses

Phenotypic and genotypic data were analyzed in RStudio (R version 3.4.4) using the packages lme4 (Bates et al., 2015) and BreedR (Muñoz and Rodriguez, 2020), and data visualization was created with ggplot2 (Wickham, 2016). The phenotypic values for the QTL analysis of the multiparent population were calculated by fitting a model with the subpopulation and year effects as fixed and the genotype effect as random. For the data from the Benning × Danbaekkong RIL population, best linear unbiased predictions (BLUPs) were obtained by fitting a model with the environment (location + year) as a fixed effect and genotype and replication as random effects. Analysis of the phenotypic data from yield trials with the 103 selected breeding lines was performed by fitting a model with the QTL within each subpopulation as a fixed effect, and genotype, environment, and replication as random effects.

To saturate the QTL region identified in the Benning × Danbaekkong RIL population, additional markers were developed in the QTL interval based on comparison of the sequencing data from the seven sequenced genotypes. Linkage map construction and QTL analysis were performed with the R package R/qtl (Broman et al., 2003). Associations between markers and protein content were established with a regression function using a LOD significance threshold determined by 1,000 permutations. Recombination distances were calculated using Kosambi's mapping function with simple interval and composite interval mapping methods in the QTL position estimation. To understand the effects of the QTL in a broad genetic background, a multiparent population QTL analysis was performed using an R package mppR (Garin et al., 2020). In each round of mapping, the population was randomly partitioned into five subsets and one of the subsets was used for validation of the parameters calculated in the other four subsets. Composite interval mapping was performed in each subset 100 times and the QTL position was determined by the location of the most significant marker across all iterations.

3 Results

3.1 Danbaekkong high-protein allele

The RIL population derived from the Benning × Danbaekkong cross (N = 140) was genotyped with the Chr 20 QTL flanking markers previously used by Warrington et al. (2015) and 17 new additional markers designed based on variants found in the comparison of the seven sequenced lines. The markers were combined to saturate the Chr 20 QTL region and one of the markers used (GSM1252) specifically targeted the 321-bp insertion in the gene *Glyma.20g085100* identified by Fliege et al. (2022). The QTL interval was identified in a genomic region between 27.7 and 33.0 Mb across all environments tested and the marker GSM1252 designed from the gene *Glyma.20g085100* was one of the most significant markers across the environments (Figure 1). In this population, the QTL explained 47.5% of the phenotypic variation and had an additive effect of 1.3% in the protein content. The homozygous RILs for the low-protein allele at the GSM1252 locus had an average protein content of 43.8%, while the lines with the homozygous high-protein allele had a protein content of 46.4%.

QTL mapping was also performed in the multiparent population and the QTL region was identified to the interval between 31.8 and 32.2 Mb (markers GSM1252 and GSM0455). This region is located within the QTL interval identified in the analysis of the Benning × Danbaekkong RIL population (Figure 2). After the estimation of the QTL parameters, an association analysis between the *Glyma.20g85100* marker (GSM1252) and the content of protein and oil was performed using the lines in the multiparent population. The high-protein allele from Danbaekkong was associated with an increase in the protein of 3.3% on average (ranging from 2.6% to 3.7%) across all populations. The highest increase in protein content was observed in population G13-6299 ×

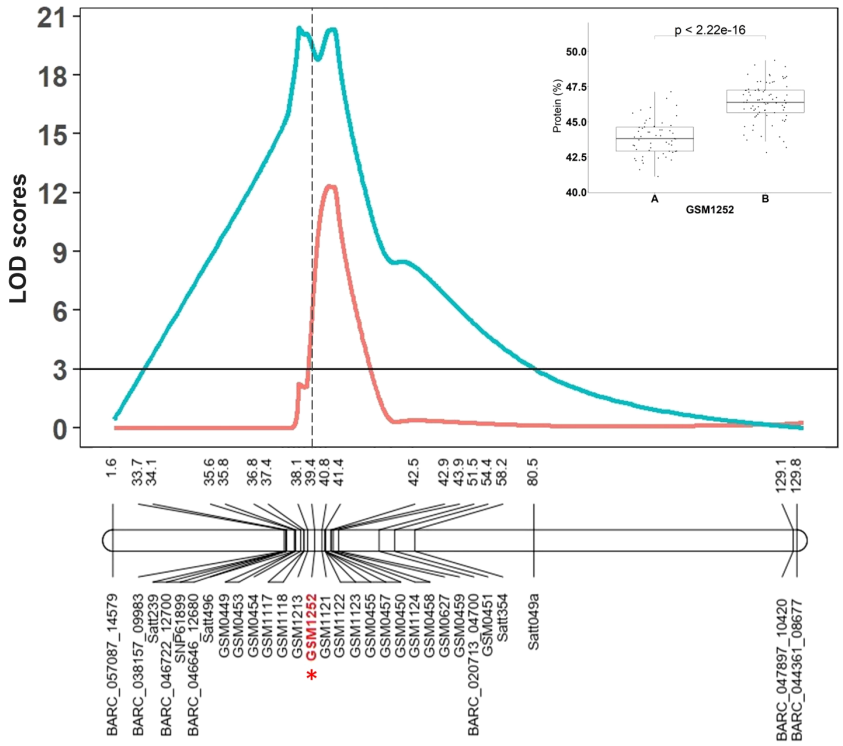


FIGURE 1
The Chr 20 QTL region identified by Warrington et al. (2015) in the RIL population derived from Benning × Danbaekkong and saturated with additional KASP markers and the gene-specific TaqMan marker GSM1252 is indicated in red with an asterisk. Red lines indicate Composite Interval Mapping and blue lines indicate Simple Interval Mapping. Marker distances are given in centimorgan (cM). Additional information about the markers is presented in Supplementary Tables S3–S5.

Benning HP, with protein content going from 40.8% to 44.5%. The highest average value of protein obtained was in the population Woodruff × Benning HP with lines carrying the high-protein allele reaching 45.4% (Figures 3, 4; Supplementary Table S6).

The increase in the protein content was accompanied by a reduction in oil content in all populations, ranging from a reduction of 1.4% in Benning HP × G10PR-56444R2 to 2.0% in N10-711 × Benning HP and Benning HP × G11PR-56238R2. On average, for every 1.8% increase in protein, there was a decrease of 1% in oil. The populations N08-174 × Benning HP and Benning HP × G10PR-56444R2 had an average oil content ≥20% and protein content ≥43.5%, demonstrating the possibility of having high protein and oil above 20% (Supplementary Table S6).

The fact that Benning HP was used either as a male or female parent in the multiparent population enabled evaluation of any maternal effect of the Danbaekkong high-protein allele. It was observed that the Danbaekkong high-protein allele increased the protein in a similar magnitude having the Benning HP as the female (44.5%) or the male (44.3%) parent in the cross (Supplementary Table S6).

3.2 Effects of the Danbaekkong high-protein allele on yield

To assess the effects of the protein QTL on yield, 103 RILs were selected from the multiparent population based on agronomic performance and visual assessment of plant appearance, lodging, and maturity to enter the 2020 and 2021 yield trials. Population N10-711 × Benning HP had the highest number of lines in the trials (27 in 2020 and 13 in 2021), while Benning HP × G11PR-56238R2 had the lowest number, with three lines in the yield trials. Overall, all pedigrees had lines with the high-protein allele or low-protein allele variant evaluated in both years, except for the Benning HP × G11PR-56238R2 population, which was evaluated only in 2020 and Benning HP × G10PR-56444R2 did not have lines with the high-protein allele tested. Having lines with and without the Danbaekkong high-protein allele evaluated in yield trials in 9 of the 10 pedigrees enabled a comparison of the effects of the increased protein content on yield in multiple genetic backgrounds.

In the yield trials, the lines carrying the high-protein allele had a consistently higher protein content across all the populations, with an average increase of 2.0% in protein content. The only exception was population R12-514 × Benning HP, in which lines with the high-protein allele in the population did not have a significant increase in protein content. Population G13-6299 × Benning HP had the highest increase in protein, from 40.1% to 43.2% and population N10-711 × Benning HP had the highest average value of protein, with 43.8% (Figure 5A). The oil content had an overall reduction of 1%, but variation was observed across the different populations, ranging from a 2% reduction in Benning HP × G11PR-56238R2 to no detectable reduction in R12-514 × Benning HP (Figure 5B). In the comparison

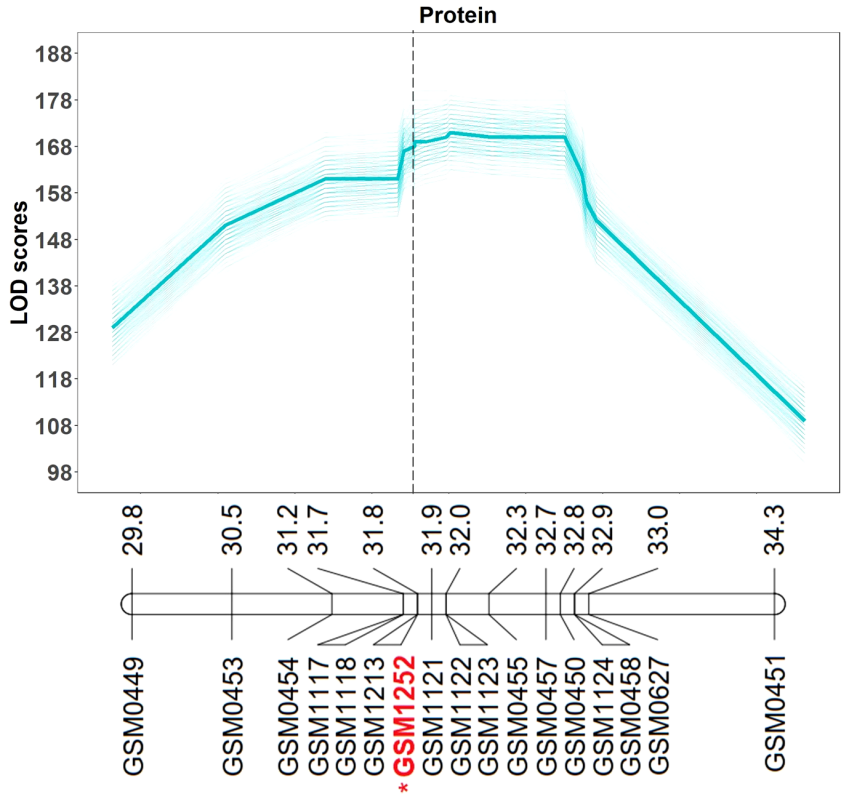


FIGURE 2
Multiparent population QTL analysis for seed protein and oil content. QTL analysis was performed 500 times (5 random subsets with 100 replications) using the composite interval mapping function. The average LOD value of all values is indicated in the bold line. Mapping was performed with the KASP markers and the gene-specific TaqMan marker GSM1252 indicated in red with an asterisk. Marker position is given in Mb based on Wm82.a2.v1. Additional information about the markers is presented in Supplementary Tables S3–S5.

of the protein production per hectare, the populations also had different performances. Lines with the high-protein allele from the population N10-711 × Benning HP had an increase of 94 kg ha^{-1} in protein production, but in the population N05-7432 × Benning HP, the lines with the high-protein allele had a decrease of 218 kg ha^{-1}. When considering the performance of all populations together, there was no difference (p = 0.41) in the protein production per hectare in the lines with or without the Danbaekkong high-protein allele, 2,048 vs. 2,080 kg ha^{-1}, respectively (Figure 5C; Supplementary Table S7).

Overall, the high protein negatively impacts the yields, with an average reduction of 313 kg ha^{-1}. However, there was variation across the different populations, with population N05-7432 × Benning HP having a yield reduction of 719 kg ha^{-1} to the population N10-711 × Benning HP with a yield reduction of only 55 kg ha^{-1} in the lines with the high-protein allele. Of the 103 lines evaluated, 20 lines from different populations had yield similar to or higher than the commercial check AGS 738RR, and 14 of these lines had a protein content higher than 40% (Figure 5D; Supplementary Table S8). The line G19-11395 from population N05-7432 × Benning HP did not have the high-protein allele but stood out with the highest overall yield, 5,880 kg ha^{-1}, 13.8% higher but not significantly different from AGS-738RR. The line G19-11191 from population Woodruff × Benning HP was the only line carrying the high-protein allele (43.6% protein) that had yield comparable to the AGS 738RR (100.4%), with 5,189 kg ha^{-1}. Three other lines, G19-11422 (N05-7432 × Benning HP), G19-11111 (G13-6299 × Benning HP), and G19-2139R2 (Benning HP × G11PR-56151R2), carrying the high-protein allele at GSM1252, had a protein content exceeding 43% and yielded >95% of AGS 738RR (Supplementary Table S8). These results exemplify the possibility of combining high yield with improved seed composition.

3.3 Effect of maturity on seed protein

The association between maturity and the high-protein alleles was evaluated across the different pedigrees in the multiparent population. Nine out of 10 populations studied had the lines carrying the high-protein allele reaching maturity earlier than those with normal protein. Overall, high-protein lines reached maturity 3.5 days earlier than those with low protein (Table 1). The population with the biggest difference was N05-7432 × Benning HP, in which the lines having the high-protein allele matured 6.1 days earlier than those with the low-protein allele. On the other hand, there was no significant difference in maturity between lines with the high-protein allele and those with the low-protein allele in the population N08-174 × Benning HP.

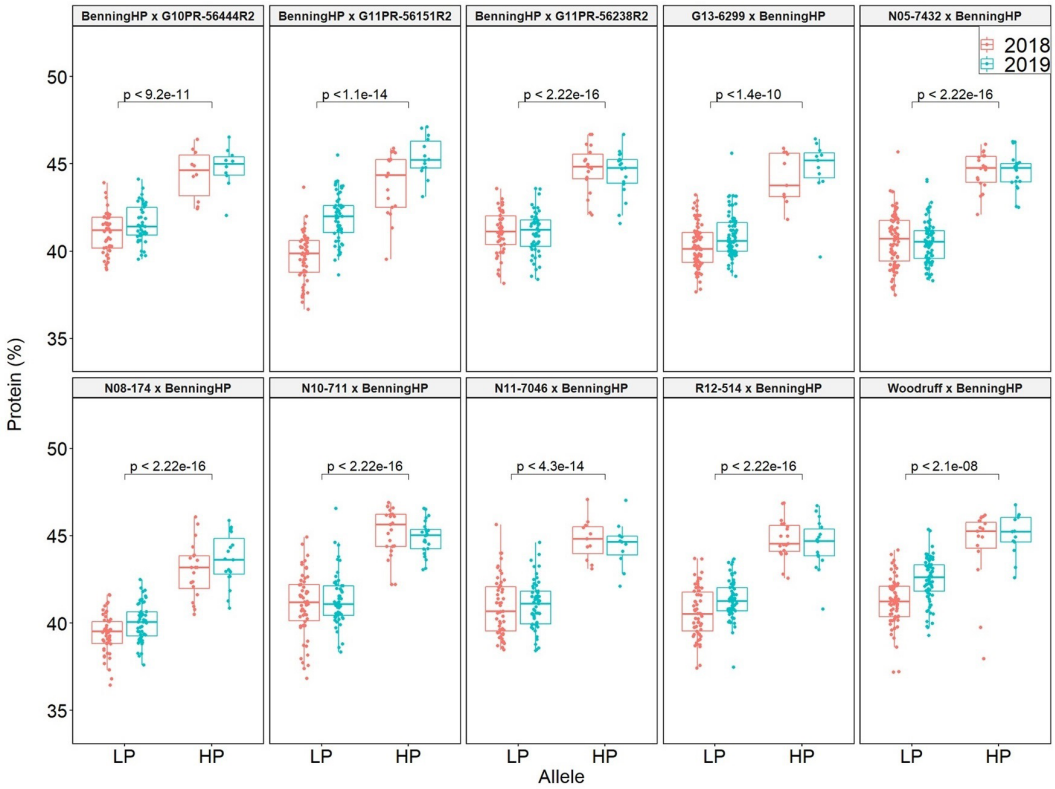


FIGURE 3
Effects of the Danbaekkong Chr 20 high-protein allele on seed protein in 10 RIL populations evaluated in 2018–2019. The *X* axis indicates the allele at the *Glyma.20g085100* (GSM1252). HP and LP represent the high- and low-protein alleles as indicated by the gene-specific marker GSM1252, respectively. Protein content is on a dry matter basis.

3.4 Distribution of the Chr 20 high-protein allele among the soybean ancestors and *G. soja* lines

The presence of the Danbaekkong high-protein allele was surveyed using the gene-specific TaqMan marker GSM1252 in a panel of 35 *G. max* ancestral lines that contributed 95% of the genes found in modern soybean cultivars (Gizlice et al., 1994). These lines provided a good opportunity to understand the distribution of the high-protein allele in the North American soybean breeding pool. The results indicated that all 35 *G. max* ancestors have the low-protein allele at the gene *Glyma.20g085100*, and the average protein content was 41.6% (ranging from 38.1% to 45.7%) (Table 2; Figure 6).

Another analysis was performed to study the distribution of the high-protein allele across *G. soja* accessions. A panel of 79 diverse *G. soja* that represent the genetic diversity in USDA Soybean Germplasm Collection was surveyed (La et al., 2019). All the *G. soja* lines evaluated presented the high-protein allele on the Chr 20 and had an average protein content of 44.4% (ranging from 39.8% to 49.4%) (Table 3; Figure 6). To confirm the presence of the high-protein allele in *G. soja*, the sequence of 35 accessions that have not been studied previously was analyzed for the presence of the insertion in *Glyma.20g085100*. Confirming the previous results, all *G. soja* lines evaluated have the high-protein allelic variant (Supplementary Table S9).

4 Discussion

4.1 Danbaekkong high-protein allele

Using new molecular markers positioned in the interval where the protein QTL has been repeatedly identified (29.8 to 34.3 Mb), genotyping was performed in the Benning × Danbaekkong RIL population (N = 140) and in a multiparent population (N = 1,115). Of these markers, GSM1252 was developed based on previous research that identified *Glyma.20g085100* controlling the protein at the Chr 20 QTL (Fliege et al., 2022; Goettel et al., 2022). GSM1252 was developed as a TaqMan marker with one probe targeting the flanking regions of the insertion aiming to capture the alleles without the 321-bp insertion and another probe that binds to a fragment of the insertion and the right flanking site (Supplementary Figure S2). Overall, the marker exhibited a good performance in separating the lines with and without the insertion and it is a useful tool to select lines for high protein. The QTL analysis confirmed the variation in *Glyma.20g085100* to be associated with protein content in the populations derived from Danbaekkong. However, instead of GSM1252, marker GSM1122 was the most significant marker at the locus. This can be attributed to the fact that Chr 20 QTL is a region of strong linkage disequilibrium (Vaughn et al., 2014). The data analysis in Benning × Danbaekkong and the multiparent population indicated a confidence interval of 503,806 bp between the markers GSM1252 and GSM0455

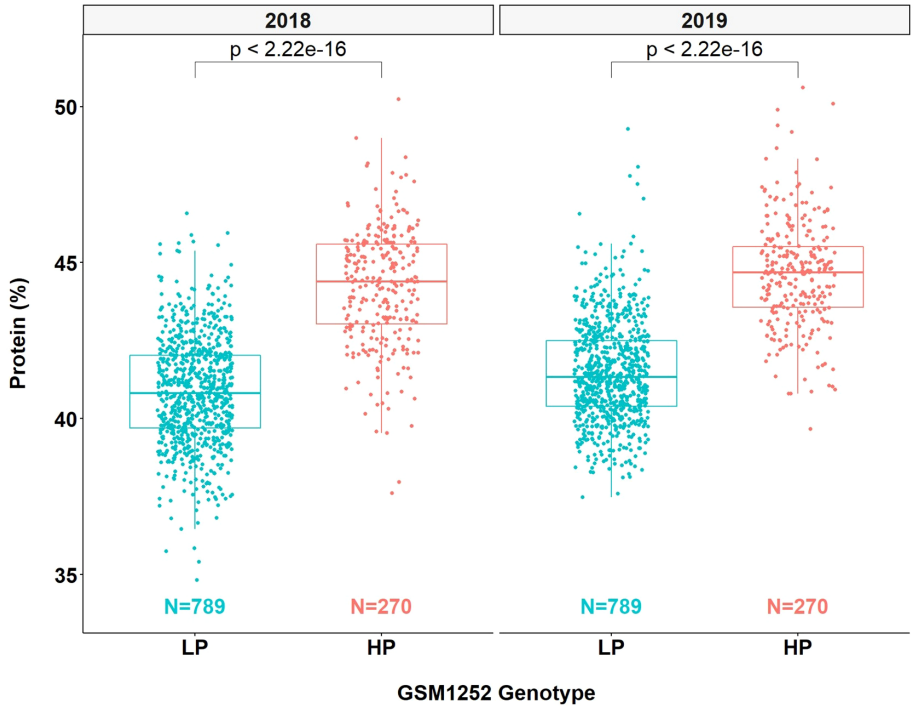


FIGURE 4
Effects of different alleles at *Glyma.20g085100* on protein content across the multiparent RIL populations. HP indicates the high-protein allele and LP indicates the low-protein allele at GSM1252.

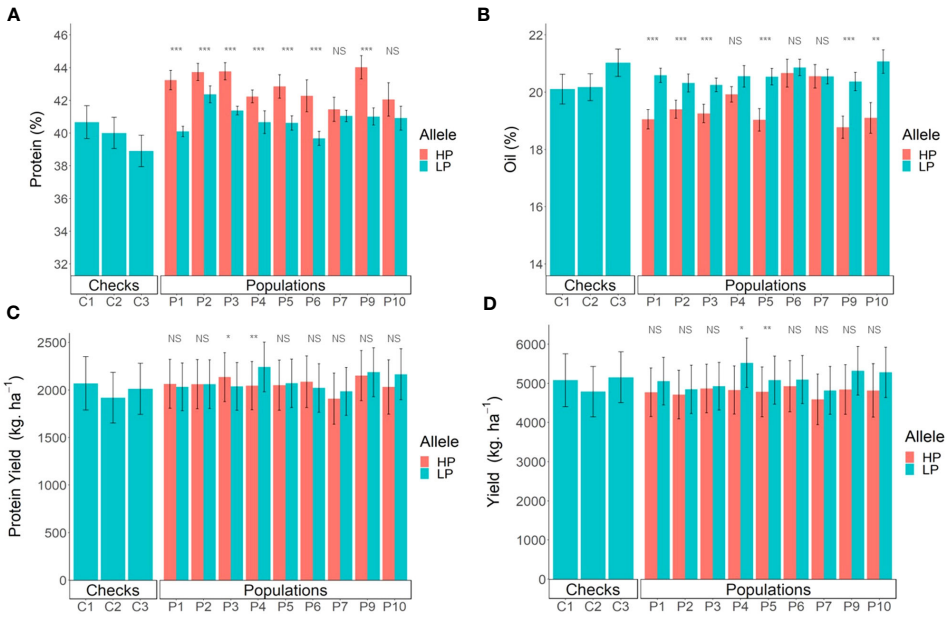


FIGURE 5
Comparison of breeding lines with and without the Danbaekkong Chr 20 high-protein allele in each population. Red indicates lines with the high-protein allele (HP) and blue indicates lines with the low-protein allele (LP). **(A)** Comparison of the protein content, **(B)** Oil content, **(C)** Production of protein per hectare, and **(D)** Seed yield. A total of 103 RILs were evaluated in five environments (Athens 2020 and 2021, Plains 2020 and 2021, and Tifton 2021) with two to three replications per environment. Error bar indicates standard error. Checks C1, C2, and C3 correspond to AG 64X8RR2X, AG 74X8RR2X, and AGS 738RR, respectively. *, **, and *** indicate significance at the 0.05, 0.01, and 0.001 probability level and NS indicates not significant.

10.3389/fpls.2023.1308731

TABLE 1 Effects of the high-protein QTL on maturity.

Population	Pedigree	Maturity†		Difference
		HP lines	LP lines	
P1	G13-6299 × Benning HP	52.8	55.3	2.5*
P2	Woodruff × Benning HP	53.0	55.5	2.5***
P3	N10-711 × Benning HP	47.8	52.3	4.5***
P4	N05-7432 × Benning HP	51.1	57.2	6.1***
P5	N11-7046 × Benning HP	48.7	52.0	3.3**
P6	N08-174 × Benning HP	41.4	41.9	0.5^{ns}
P7	R12-514 × Benning HP	40.7	44.2	3.4*
P8	Benning HP × G10PR-56444R2	52.4	55.3	2.9***
P9	Benning HP × G11PR-56151R2	48.7	54.3	5.6***
P10	Benning HP × G11PR-56238R2	48.4	53.5	5.1***

HP indicates the high-protein allele and LP indicates the low-protein allele at GSM1252.
*, **, and *** indicated significant differences at the 0.05, 0.01, and 0.001 probability level, respectively.
† Maturity is indicated as days after August 31.

TABLE 2 Distribution of the low-protein allele (321-bp insertion) among North American soybean ancestral lines as defined by Gizlice et al. (1994).

	ID	Origin	MG	Protein (%)	Oil (%)	GSM1252†
1	FC 31745	–	VI	40.2	21.5	LP
2	FC 33243	–	IV	38.1	22.5	LP
3	PI 180501‡	Germany	0	39.1	21.3	LP
4	PI 240664‡	Philippines	X	44.8	21.1	LP
5	PI 360955B‡	Sweden	0	42.7	18.2	LP
6	PI 438471	Sweden	0	38.2	20.3	LP
7	PI 438477	Sweden	0	39.6	19.7	LP
8	PI 548298	China	III	43.0	19.9	LP
9	PI 548302	Japan	II	42.2	17.8	LP
10	PI 548311‡	Canada	0	42.0	20.4	LP
11	PI 548318‡	China	III	39.1	21.6	LP
12	PI 548325	Russia	0	41.5	19.7	LP
13	PI 548348	China	III	41.5	20.0	LP
14	PI 548352‡	North Korea	III	41.4	19	LP
15	PI 548356‡	North Korea	II	41.4	19.9	LP
16	PI 548360	North Korea	II	39.7	21.4	LP
17	PI 548362	United States	III	38.4	22.9	LP
18	PI 548379	China	0	38.4	20.9	LP
19	PI 548382‡	–	0	43.1	17.6	LP
20	PI 548391	China	II	43.0	20.3	LP
21	PI 548402‡	China	IV	38.2	18.5	LP
22	PI 548406	China	II	41.6	19.0	LP

(Continued)

TABLE 2 Continued

	ID	Origin	MG	Protein (%)	Oil (%)	GSM1252†
23	PI 548438	North Korea	VI	44.7	19.2	LP
24	PI 548445	China	VII	45.7	19.0	LP
25	PI 548456	North Korea	VI	41.0	19.1	LP
26	PI 548461	United States	VIII	40.5	22.5	LP
27	PI 548477	United States	VI	42.9	20.2	LP
28	PI 548484	North Korea	VI	42.1	20.2	LP
29	PI 548485	China	VII	42.1	20.7	LP
30	PI 548488	China	V	43.8	18.9	LP
31	PI 548603	United States	IV	40.5	21.9	LP
32	PI 548657	United States	VII	40.3	21.9	LP
33	PI 71506	China	IV	41.0	22.6	LP
34	PI 80837‡	Japan	IV	42.4	18.2	LP
35	PI 88788	China	III	43.4	15.7	LP
	Benning §	United States	VII	41.9	21.3	LP
	Benning HP §	United States	VII	45.6	19.0	HP
	Danbaekkong¶	South Korea	V	48.0	18.5	HP

Protein and oil analyzed with near-infrared (NIR) spectroscopy using a sample of approximately 200 seeds harvested in 2018.
Benning, Benning HP, and Danbaekkong are controls.
† GSM1252 indicates the presence of the high-protein allele (HP) or the low-protein allele (LP).
‡ Protein and oil content retrieved from GRIN. https://npgsweb.ars-grin.gov/gringlobal/.
§ Benning and Benning HP values are averages from 3 years of tests (2019, 2020, and 2021).
¶ Danbaekkong value is the average from 2 years of tests (2017 and 2021).

(31,778,817–32,282,623 bp) (Wm82.a2.v1). This region overlaps perfectly with previously published mapping work that identified the Chr 20 QTL (Bolon et al., 2010; Vaughn et al., 2014; Warrington et al., 2014; Lee et al., 2019; Wang et al., 2021). In the analysis of the multiparent mapping population, the flat QTL peak in the region between 31.8 and 32.8 Mb indicated that this genomic region has a large linkage disequilibrium block.

To elucidate the origins of the Danbaekkong high-protein allele, an analysis of the Danbaekkong pedigree was conducted. One of the Danbaekkong's parents is the cultivar Dongsan 69 from South Korea and the pedigree of this cultivar is unknown since no release information is available. The other parent is D76-8070, which is an MG V line developed by Edgar Hartwig in his effort to breed soybean cultivars with increased protein content (Hartwig, 1990). D76-8070 was developed through the selection of progeny from multiple crosses ("Hill" × "Sioux", FC 31745 × D49-2510, Hill × PI 96983, and D49-24914 × PI 163453). The progeny from each of these crosses were selected for disease resistance, agronomic traits, and high protein content (>45%) and the selected lines were intercrossed to develop D76-8070 (Supplementary Figure S3). PI 163453 is the only *G. soja* line present in the pedigree of D76-8070 and was hypothesized as the origin for the high-protein QTL. To verify this hypothesis, the haplotypes of PI 163453 and Danbaekkong were compared using the 6,353 SNPs between 30 and 34 Mb on Chr 20 called from the sequencing data. The genetic similarity analysis showed that the Danbaekkong haplotype at the Chr 20 QTL region is 99.95% identical to PI 163453 (Supplementary Table S10). To confirm the inheritance of the protein QTL, D76-8070 was also genotyped with GSM1252 and the results indicated that it carries the same allele as PI 163453, Danbaekkong, and Benning HP (Supplementary Table S11).

To quantify the Chr 20 fragment that was transferred from PI 163453 to Danbaekkong, 408 SNPs from the SoySNP50K SNP dataset distributed along the Chr 20 were used and it was observed that the PI 163453 fragment that was transferred to D76-8070 spans from 21 to 34.6 Mb and the D76-8070 fragment that was transferred to Danbaekkong starts at 2 Mb and ends at 36 Mb. Subsequently, a fragment from 0.2 to 37 Mb from Danbaekkong was transferred to Benning HP (Supplementary Figure S4). These results indicate that the high-protein allele is originally from PI 163453, and it was transferred to D76-8070 through the work of Hartwig. Then, D76-8070 was used in South Korea to develop Danbaekkong, which eventually returned to the United States and was used to develop the isogenic line Benning HP.

The haplotype of PI 163453 was also compared to the *G. soja* line PI 468916 used in the mapping study that identified the Chr 20 QTL (Diers et al., 1992). The comparison revealed that PI 163453 is only 43% similar to PI 468916 when considering all the SNPs in the 30–34 Mb window, but when comparing the sequence of the gene *Glyma.20g085100*, it was observed that PI 163463 is also missing the

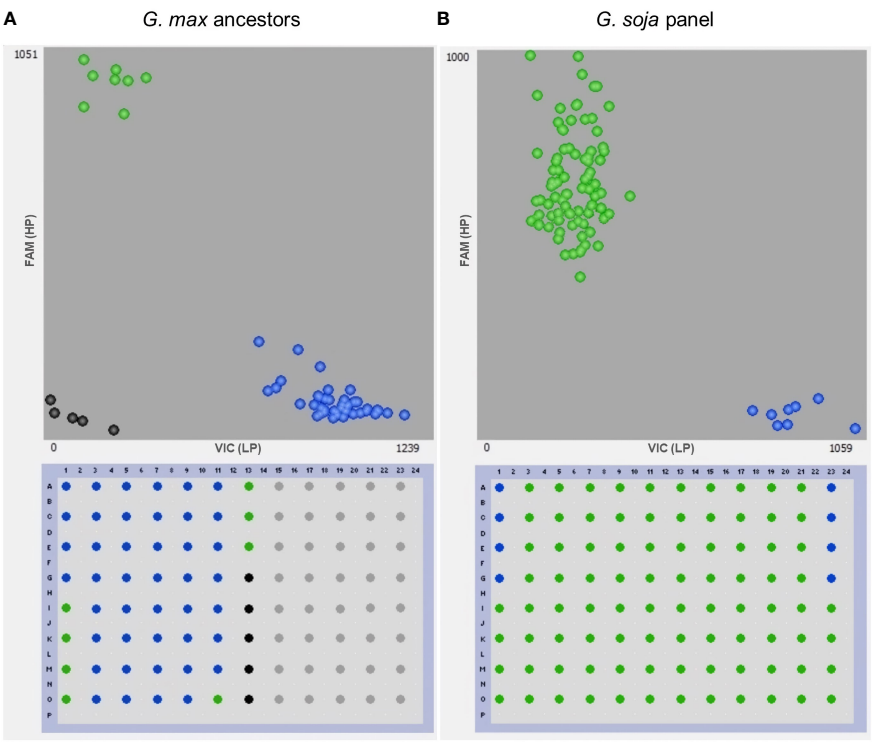


FIGURE 6
Genotyping with the gene-specific TaqMan marker GSM1252. **(A)** 35 North America *Glycine max* ancestors and **(B)** 79 *Glycine soja* accessions. Benning and Benning HP were used as controls four times on each plate for the low-protein (LP) and the high-protein allele (HP), respectively. Points in green, blue, and black colors denote the high-protein allele, low-protein allele, and no template control, respectively.

TABLE 3 Distribution of the high-protein allele among the USDA *Glycine soja* core set as defined by La et al. (2019).

	Name	Origin	MG	Protein (%)	Oil (%)	GSM1252[†]
1	PI 101404A	China	II	45.7	16.2	HP
2	PI 339871A	South Korea	V	42.9	16.6	HP
3	PI 342622A	Russia	I	43.7	16.1	HP
4	PI 366122	Japan	IV	44.1	16.6	HP
5	PI 378683	Japan	VI	46.7	16.4	HP
6	PI 378684B	Japan	VI	47.3	16.0	HP
7	PI 378686B	Japan	VI	46.0	16.3	HP
8	PI 378690	Japan	VII	45.3	16.3	HP
9	PI 378696B	Japan	VI	43.7	16.7	HP
10	PI 378697A	Japan	V	44.5	16.5	HP
11	PI 407020	Japan	V	44.0	16.8	HP
12	PI 407038	Japan	V	45.4	16.5	HP
13	PI 407042	Japan	V	44.9	16.3	HP
14	PI 407052	Japan	V	46.8	16.1	HP
15	PI 407059	Japan	–	46.7	16.1	HP
16	PI 407085	Japan	VI	44.8	16.5	HP

(Continued)

10.3389/fpls.2023.1308731

TABLE 3 Continued

	Name	Origin	MG	Protein (%)	Oil (%)	GSM1252[†]
17	PI 407096	Japan	VII	47.2	16.3	HP
18	PI 407156	Japan	VI	44.7	16.5	HP
19	PI 407157	Japan	VI	47.8	16.3	HP
20	PI 407171	South Korea	IV	43.8	16.4	HP
21	PI 407179	South Korea	V	44.4	16.8	HP
22	PI 407191	South Korea	V	46.2	16.5	HP
23	PI 407195	South Korea	IV	44.4	16.5	HP
24	PI 407206	South Korea	V	46.4	16.3	HP
25	PI 407214	South Korea	V	46.7	16.4	HP
26	PI 407228	South Korea	V	49.5	15.8	HP
27	PI 407231	South Korea	V	44.4	16.5	HP
28	PI 407240	South Korea	V	46.3	16.5	HP
29	PI 407248	South Korea	V	44.6	16.6	HP
30	PI 407287	Japan	V	45.6	16.3	HP
31	PI 407300	China	V	46.1	16.2	HP
32	PI 407314	South Korea	V	44.2	16.9	HP
33	PI 424004B	South Korea	II	43.6	16.5	HP
34	PI 424007	South Korea	V	42.3	16.8	HP
35	PI 424025B	South Korea	V	46.3	16.4	HP
36	PI 424035	South Korea	V	43.3	16.7	HP
37	PI 424045	South Korea	V	42.6	16.5	HP
38	PI 424070B	South Korea	V	43.3	16.5	HP
39	PI 424082	South Korea	V	44.1	16.1	HP
40	PI 424083A	South Korea	V	45.4	16.4	HP
41	PI 424102A	South Korea	V	43.6	16.5	HP
42	PI 424116	South Korea	IV	43.7	16.6	HP
43	PI 424123	South Korea	V	44.0	16.1	HP
44	PI 447003A	China	0	43.8	16.8	HP
45	PI 458536	China	0	48.3	16.3	HP
46	PI 464890B	China	I	47.2	16.2	HP
47	PI 479746B	China	II	46.6	16.1	HP
48	PI 479751	China	III	43.7	16.8	HP
49	PI 479752	China	I	41.2	16.4	HP
50	PI 479768	China	0	44.8	16.4	HP
51	PI 483466	China	V	43.9	16.2	HP
52	PI 507618	Japan	V	44.1	16.4	HP
53	PI 507624	Japan	VII	44.6	16.4	HP
54	PI 507641	Japan	V	45.9	16.6	HP
55	PI 507656	Japan	VII	45.9	16.3	HP

(Continued)

TABLE 3 Continued

	Name	Origin	MG	Protein (%)	Oil (%)	GSM1252†
56	PI 507761	Russia	I	42.4	16.4	HP
57	PI 522209B	Russia	II	43.2	16.4	HP
58	PI 522226	Russia	000	43.3	16.3	HP
59	PI 522233	Russia	I	44.3	16.1	HP
60	PI 522235B	Russia	I	41.6	16.2	HP
61	PI 549032	China	III	44.0	15.9	HP
62	PI 549046	China	III	39.9	17.1	HP
63	PI 549048	China	III	41.0	17.6	HP
64	PI 562547	South Korea	V	41.2	16.5	HP
65	PI 562551	South Korea	V	43.9	16.5	HP
66	PI 562553	South Korea	V	47.4	16.3	HP
67	PI 562561	South Korea	V	47.1	16.0	HP
68	PI 562565	South Korea	IV	43.2	16.4	HP
69	PI 593983	Japan	III	45.0	16.7	HP
70	PI 597448D	China	0	45.2	16.2	HP
71	PI 597458C	China	V	43.5	17.2	HP
72	PI 597460A	China	IV	42.9	16.8	HP
73	PI 597461B	China	V	39.8	17.5	HP
74	PI 597462B	China	IV	42.5	17.1	HP
75	PI 639586	Russia	–	42.0	17.0	HP
76	PI 639588B	Russia	–	41.8	17.1	HP
77	PI 639621	Russia	–	41.8	17.1	HP
78	PI 639623A	Russia	–	44.2	16.5	HP
79	PI 639635	Russia	–	43.3	16.4	HP
	PI 163453	China	VI	44.7	12.0	HP
	PI 468916	China	III	44.0	10.1	HP
	Benning‡	United States	VII	41.9	21.3	LP
	Benning HP‡	United States	VII	45.6	19.0	HP
	Danbaekkong§	South Korea	V	48.0	18.5	HP

Protein and oil contents for *G. soja* accessions were obtained from La et al. (2019).
All *G. soja* accessions have black seed coat color.
Benning, Benning HP, and Danbaekkong are controls. PI 163453 is the *G. soja* ancestor of Danbaekkong and PI 468916 is the *G. soja* used in Fliege et al. (2022).
† GSM1252 indicates the presence of the high-protein allele (HP) or the low-protein allele (LP).
‡ Benning and Benning HP values are averages from 3 years of tests (2019, 2020, and 2021).
§ Danbaekkong value is the average from 2 years of tests (2017 and 2021).

321-bp fragment as PI 468916 (Supplementary Table S11, Supplementary Figure S5). These results indicate that although PI 163453 and PI 468916 are different at the haplotype level, they carry the same high-protein allele in *Glyma.20g085100*.

Goettel et al. (2022) indicated that the *Glyma.20g085100* high-protein allele was transferred from *G. soja* to *G. max* in three independent events likely during the process of domestication in East Asia. In the present research, it was demonstrated that the Danbaekkong high-protein allele came from the intentional introgression conducted by Edgar Hartwig where the *G. soja* PI 163453 was used as a grand parent to develop D76-8070 (Hartwig, 1990). Analyzing the haplotypes in the *Glyma.20g085100* region (Chr20, 29–34 Mb) revealed that both PI 163453 and Danbaekkong were grouped into cluster 3 identified by Goettel et al. (2022) (Figure 7). Cluster 3 is predominantly composed of the accessions from China except Danbaekkong that is a derived progeny from PI 163453.

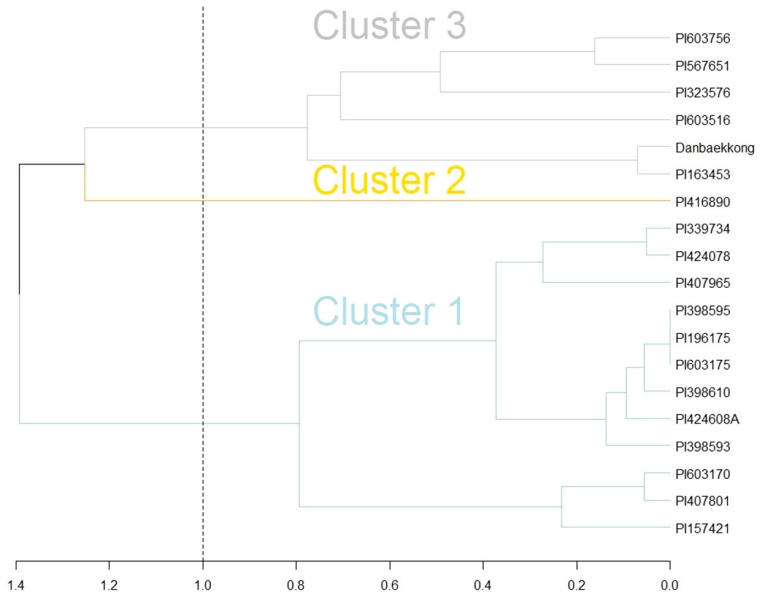


FIGURE 7
Comparison of PI 163453 and Danbaekkong haplotypes with the three introgression groups identified by Goettel et al. (2022) using hierarchical complete linkage cluster analysis. Analysis was based on 82 SNPs from the SoySNP50K at the Chr 20 QTL region between 29 and 34 Mb.

4.2 Distribution of the Chr 20 high-protein allele among the soybean ancestors and *G. soja* lines

An analysis of the distribution of the high-protein allele was performed using 35 *G. max* that represents the diversity of the North American soybean cultivars (Gizlice et al., 1994). The results indicated that none of the 35 *G. max* ancestors carry the high-protein allele in *Glyma.20g085100*. However, three soybean ancestors, CNS (PI 548445), Arksoy (PI 548438), and Bilomi No. 3 (PI 240664), have protein content higher than 44% but do not carry the Chr 20 high-protein allele. CNS, Arksoy, and Bilomi No. 3 were originally collected in China, North Korea, and Philippines, respectively, and it is possible that these three accessions harbor protein QTLs in other genomic regions. To our knowledge, these ancestors have not been used in QTL mapping studies yet and they could reveal more information about the genetic control of protein in soybean.

Soybean lines with protein content reaching values of 47.2% have been developed (Wilcox and Cavins, 1995), and some lines have been released as cultivars in the United States in an effort to improve the seed composition, such as Protana with 43% protein (Probst et al., 1971), Prolina with 46% protein (Burton et al., 1999), and Prohio with 44.1% protein (Mian et al., 2008). More recently, soybean breeders focused on combining high yield and improved protein content and several breeding lines have been released. Chen et al. (2017) developed UA 5814HP as a new soybean cultivar with high seed protein content (45.5%) and yield comparable to elite checks. Pantalone and Smallwood (2018) released TN11-5102 as a high-yield and high-protein line with 42% protein. Shannon et al. (2022) developed S09-13185, with 44% protein content and Li et al. (2022) released G11-7013 with a protein content of 43.6%. Despite these efforts, the proportion of high-protein lines in North American germplasm is low. According to Patil et al. (2017), most soybean cultivars in the United States are fixed for the low-protein allele at the Chr 20 locus, and the introgression of the high-protein allele has the potential to improve the seed protein content in soybean cultivars in North America.

Goettel et al. (2022) analyzed a panel of 398 *G. max* (259 Cultivars and 139 Landraces) and 150 *G. soja* accessions from the USDA Soybean Germplasm Collection and observed that only 21 *G. max* lines had the Chr 20 high-protein allele. Of these 21 *G. max* lines that have the high-protein allele, 1 line was from India, 2 lines were from Japan, 4 lines were from China, and 14 lines were from South Korea, where Danbaekkong originated. Eight of the 14 Korean lines are cultivars with yellow seed coat, indicating that the Chr 20 high-protein allele has been selected and used in the development of soybean cultivars in Korean breeding programs. Lee et al. (2015) conducted a pedigree reconstruction of Korean soybean varieties and demonstrated that since 1913, soybean breeding programs have focused primarily on the improvement of seed protein composition for processing as soy food, such as soy sauce and tofu.

Differently from *G. max*, it was observed that all 79 *G. soja* from the USDA core collection analyzed carry the high-protein allele at *Glyma.20g085100*. When analyzing the sequence of additional 35 *G. soja* accessions, all of them also carry the high-protein allele. In a similar way, Goettel et al. (2022) analyzed a panel of 150 *G. soja* accessions and found that 147 lines had the high-protein allele confirmed. Owing to the widespread presence in *G. soja* of the high-protein allele in *Glyma.20g085100* and the low frequency in *G. max*, and the fact that *G. soja* is the closest ancestor to *G. max*, it is possible to infer that the high-protein allele is the original state of the gene. A few *G. soja* accessions present a low protein content,

despite having the high-protein allele, especially the accessions PI 549046 and PI 597461B that showed a protein content lower than 40%. In fact, other studies have shown that on very few occasions, lines with the high-protein allele might have a low protein content, such as PI 407877B, PI 423954, and PI 424148 in Fliege et al. (2022). Goettel et al. (2022) indicated few wild soybean lines with a lower protein content but the overall mean protein of the *G. soja* with the high-protein allele was higher than the *G. max* with the low-protein allele. Vaughn et al. (2014) have also observed some cases where lines with the high-protein haplotype present a relatively low protein content. This is not fully understood; however, Kim et al. (2023) suggested that the protein content could be regulated by the interaction of multiple genes located at approximately 30 Mb on chromosome 20. Despite this, *Glyma.20g085100*, which is the gene targeted in this study, is likely the major gene in this regulation. The Chr 20 QTL has been shown to explain up to 55% of the variation (Warrington et al., 2015). Since it is not 100%, soybean genotypes can have relatively high or low protein through background segregation of these polygenic effects.

4.3 Effects of the Danbaekkong high-protein allele

A single marker analysis with the *Glyma.20g085100* marker was performed to understand the stability and effect of the gene across different genetic backgrounds. The analysis revealed that the high-protein allele inherited from Danbaekkong increased the protein by 3.3% on average (ranging from 2.6% to 3.7%) across all 10 populations tested in 2018 and 2019. The increase in protein content was also observed in the yield trials conducted in 2020 and 2021. In these trials, the high-protein allele had an average increase of 2.0% in the protein and only the population R12-514 × Benning HP did not show a significant increase in protein. This protein increase is similar to the estimate by Brzostowski et al. (2017), when the introgression of the Danbaekkong allele into two soybean lines caused an increase of 2% across four environments. The present results are close to the estimates by Warrington et al. (2015), where the author indicated a gain of 2.7% in protein with the Danbaekkong allele.

One of the well-known effects of the increase of protein content is the reduction of oil (Cober and Voldeng, 2000; Chung et al., 2003; Vaughn et al., 2014; Patil et al., 2018). According to Hanson et al. (1961), this relationship is dictated by a ratio of 2:1, in which the energy demanded to synthesize 2 protein units corresponds to 1 unit of oil. Other studies have shown that the protein-to-oil ratio is between 1.5 and 1.7 (Hartwig and Kilen, 1991; Chung et al., 2003). In the present research, it was observed that for every 1% increase in protein, there was a decrease of 0.55% in oil, representing a ratio of 1.8:1.

Several studies have indicated a negative relationship between protein and yield, with correlation values reaching up to −0.62 (Cober and Voldeng, 2000; Sebolt et al., 2000; Cunicelli et al., 2019). Overall, a negative correlation between these two traits appears to be common, but contrary to the omnipresent antagonist relationship between protein and oil, protein and yield do not have a consistent correlation when comparing multiple environments (Wilcox and Cavins, 1995; Prenger et al., 2019). In the present research, lines with the high-protein allele in general yield 313 kg ha^{-1} less (55 to 719 kg ha^{-1}) than those with the low-protein allele within the same population. Brzostowski et al. (2017) found a yield reduction ranging from −273 to −558 kg ha^{-1} when introgressing the Danbaekkong allele into two soybean lines. In the same way, Goettel et al. (2022) indicated that the low-protein allele at *Glyma.20g085100* is associated with a yield increase of 150.3 kg ha^{-1}. Despite the negative effect of the high-protein allele on yield, it was possible to identify lines carrying the high-protein allele (>43% protein) with comparable yield to the commercial checks (>95% yield). This shows that there is potential to couple high yield and high protein content with selection during breeding, and the negative association between protein and yield can be minimized.

An association between the presence of the high-protein allele in *Glyma.20g085100* and maturity was observed across different populations, where lines with the high-protein allele matured approximately 3.7 days earlier than their counterparts in the same population. Similar results were found by Prenger et al. (2019), where lines carrying the Danbaekkong allele matured earlier than those without the allele. The gene *Glyma.20g085100* is located 1.4 Mb upstream of the maturity locus E4 (Liu et al., 2008). Since Danbaekkong is an MG V cultivar, it is possible that it possesses the early maturity allele at the *E4* locus linked with the high-protein allele in a coupling phase. Therefore, the difference in maturity in lines with high protein derived from Danbaekkong is due to linkage between the high-protein QTL and the maturity gene *E4*.

To our knowledge, the present study was the first time a QTL for protein content in soybean has been fully assessed in a wide variety of genetic background simultaneously with several environments of yield trials, and its breeding history from *G. soja* to *G. max* has been described. This study complements and validates the findings of previous research about the role of *Glyma.20g085100* in determining the protein content in soybeans, providing more information about the effects and stability of the QTL, and confirming the value of its use to improve soybean seed composition.

5 Conclusions

In this research, a gene-specific marker, GSM1252, was designed for *Glyma.20g085100* and genotyping the bi-parental and multiparental populations confirmed the effectiveness of this marker as well as other flanking markers. This information can be useful resources for breeding programs to introgress the high-protein allele into elite lines. The analysis of the distribution of the *Glyma.20g085100* alleles revealed that the 35 *G. max* accessions that represent the genetic diversity of North American soybean cultivars have the low-protein allele, while the 79 *G. soja* accessions surveyed possess the high-protein allele. The analysis of the pedigree of Danbaekkong indicated that its high-protein allele was inherited from *G. soja* PI 163453, which is the same as the one from PI 468916. The Danbaekkong high-protein allele increased the protein content in all populations tested in 2018

and 2019 with an average of 3.3%, ranging from 2.6% in Benning HP × G10PR-56444R2 to a 3.7% increase in G13-6299 × Benning HP. The yield trials in 2020 and 2021, the allele increased the protein in 2% on average and was stable across multiple environments. It was observed that the increase in protein was accompanied by an overall decrease in oil and yield. However, it was possible to select breeding lines with the high-protein allele and yield comparable to elite checks, and this will enable the development of new cultivars with high protein content and high yield.

Data availability statement

The genome sequencing data supporting the conclusions of this study have been deposited under the NCBI SRA database under BioProject PRJNA1031345.

Author contributions

RS: Formal analysis, Data curation, Visualization, Writing – original draft. MARM: Investigation, Methodology, Resources, Writing – review & editing. JNV: Formal analysis, Software, Writing – review & editing. ZL: Conceptualization, Project administration, Methodology, Resources, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Support of this research was provided by the United Soybean Board, the University of Georgia Research Foundation, and Georgia Agricultural Experiment Stations. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)– Finance Code 001 by providing a scholarship to the first author.

Acknowledgments

We thank Tatyana Nienow, Nicole Bachleda, Dale Wood, Brice Wilson, and Brian Little at the University of Georgia for the technical support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1308731/full#supplementary-material

References

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome* 8, 1–13. doi: 10.3835/plantgenome2015.04.0024

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67, 1–48. doi: 10.18637/jss.v067.i01

Bayer, P. E., Yuan, Y., Batley, J., Nguyen, H. T., Valliyodan, B., Varshney, R. K., et al. (2021). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome* 15, 1–12. doi: 10.1002/tpg2.20109

Boerma, H. R., Hussey, R. S., Phillips, D. V., Wood, E. D., Rowan, G. B., and Finnerty, S. L. (1997). Registration of 'Benning' Soybean. *Crop Sci.* 37, 1982–1982. doi: 10.2135/cropsci1997.0011183x003700060061x

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bolon, Y. T., Joseph, B., Cannon, S. B., Graham, M. A., Diers, B. W., Farmer, A. D., et al. (2010). Complementary genetic and genomic approaches help characterize the linkage group I seed protein QTL in soybean. *BMC Plant Biol.* 10, 1–24. doi: 10.1186/1471-2229-10-41

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Brumm, T. J., and Hurburgh, C. R. (1990). Estimating the processed value of soybeans. *J. Am. Oil Chem. Soc* 67, 302–307. doi: 10.1007/BF02539680

Bryant, C. (2020). *Soybean production in Georgia. 1st ed* (Athens: University of Georgia Cooperative Extension).

Brzostowski, L. F., Pruski, T. I., Specht, J. E., and Diers, B. W. (2017). Impact of seed protein alleles from three soybean sources on seed composition and agronomic traits. *Theor. Appl. Genet.* 130, 2315–2326. doi: 10.1007/s00122-017-2961-x

Burton, J. W., Carter, T. E., and Wilson, R. F. (1999). Registration of ` prolina ' Soybean. *Crop Sci.* 39, 1993–1994. doi: 10.2135/cropsci1999.0011183X003900010066x

Chen, P., Florez-Palacios, L., Orazaly, M., Manjarrez-Sandoval, P., Wu, C., Rupe, J. C., et al. (2017). Registration of 'UA 5814HP' Soybean with high yield and high seed-protein content. *J. Plant Regist.* 11, 116–120. doi: 10.3198/jpr2016.09.0046crc

Chung, J., Babka, H. L., Graef, G. L., Staswick, P. E., Lee, D. J., Cregan, P. B., et al. (2003). The seed protein, oil, and yield QTL on soybean linkage group I. *Crop Sci.* 43, 1053–1067. doi: 10.2135/cropsci2003.1053

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/cam.20753

Cober, E. R., and Voldeng, H. D. (2000). Developing high-protein, high-yield soybean populations and lines. *Crop Sci.* 40, 39–42. doi: 10.2135/cropsci2000.40139x

Cunicelli, M. J., Bhandari, H. S., Chen, P., Sams, C. E., Mian, M. A. R., Mozzoni, L. A., et al. (2019). Effect of a mutant danbaekkong allele on soybean seed yield, protein, and oil concentration. *J. Am. Oil Chem. Soc* 96, 927–935. doi: 10.1002/aocs.12261

de Borja Reis, A. F., Tamagno, S., Moro Rosso, L. H., Ortez, O. A., Naeve, S., and Ciampitti, I. A. (2020). Historical trend on seed amino acid concentration does not follow protein changes in soybeans. *Sci. Rep.* 10, 1–10. doi: 10.1038/s41598-020-74734-1

Diers, B. W., Keim, P., Fehr, W. R., and Shoemaker, R. C. (1992). RFLP analysis of soybean seed protein and oil content. *Theor. Appl. Genet.* 83, 608–612. doi: 10.1007/BF00226905

Edwards, K., Johnstone, C., and Thompson, C. (1991). A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res.* 19, 1349. doi: 10.1093/nar/19.6.1349

Fliege, C. E., Ward, R. A., Vogel, P., Nguyen, H., Quach, T., Guo, M., et al. (2022). Fine mapping and cloning of the major seed protein quantitative trait loci on soybean chromosome 20. *Plant J.* 110, 1–15. doi: 10.1111/tpj.15658

Garin, V., Malosetti, M., and van Eeuwijk, F. (2020). Multi-parent multi-environment QTL analysis: an illustration with the EU-NAM Flint population. *Theor. Appl. Genet.* 133, 2627–2638. doi: 10.1007/s00122-020-03621-0

Gizlice, Z., Carter, T. E., and Burton, J. W. (1994). Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* 34, 1143–1151. doi: 10.2135/cropsci1994.0011183X003400050001x

Goettel, W., Zhang, H., Li, Y., Qiao, Z., Jiang, H., Hou, D., et al. (2022). POWR1 is a domestication gene pleiotropically regulating seed quality and yield in soybean. *Nat. Commun.* 13, 3051. doi: 10.1038/s41467-022-30314-7

Hanson, W. D., Leffel, R. C., and Howell, R. W. (1961). Genetic analysis of energy production in the Soybean. *Crop Sci.* 1, 121–126. doi: 10.2135/cropsci1961.0011183X000100020011x

Hartwig, E. E. (1990). Registration of soybean high-protein germplasm line 'D76-8070.'. *Crop Sci.* 30, 764–765. doi: 10.2135/cropsci1990.0011183x003000030092x

Hartwig, E. E., and Kilen, T. C. (1991). Yield and composition of soybean seed from parents with different protein, similar yield. *Crop Sci.* 31, 290–292. doi: 10.2135/cropsci1991.0011183x003100020011x

Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15, 1–12. doi: 10.1186/1471-2164-15-1

Keim, P., Olson, T. C., and Shoemaker, R. C. (1988). A rapid protocol for isolating soybean DNA. *Soybean Genet. Newsl.* 15, 150–152.

Kim, S. D., Hong, E.-H., Kim, Y.-H., Lee, S.-H., Seong, Y.-K., Park, K.-Y., et al. (1996). A new high protein and good seed quality soybean variety "Danbaegkong". *RDA. J. Agri. Sci.* 38, 228–232.

Kim, W. J., Kang, B. H., Moon, C. Y., Kang, S., Shin, S., Chowdhury, S., et al. (2023). Quantitative trait loci (QTL) analysis of seed protein and oil content in wild soybean (Glycine soja). *Int. J. Mol. Sci.* 24 (4), 4077. doi: 10.3390/ijms24044077

La, T., Large, E., Taliercio, E., Song, Q., Gillman, J. D., Xu, D., et al. (2019). Characterization of select wild soybean accessions in the USDA germplasm collection for seed composition and agronomic traits. *Crop Sci.* 59, 233–251. doi: 10.2135/cropsci2017.08.0514

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lee, C., Choi, M., Kim, H., Yun, H., Lee, B., Chung, Y., et al. (2015). Soybean [Glycine max (L.) merrill]: importance as A crop and pedigree reconstruction of korean varieties. *Plant Breed. Biotech.* 3, 179–196. doi: 10.1016/S0828-282X(08)70684-6

Lee, S., Van, K., Sung, M., Nelson, R., LaMantia, J., McHale, L. K., et al. (2019). Genome-wide association study of seed protein, oil and amino acid contents in soybean from maturity groups I to IV. *Theor. Appl. Genet.* 132, 1639–1659. doi: 10.1007/s00122-019-03304-5

Lestari, P., Van, K., Lee, J., Kang, Y. J., and Lee, S.-H. (2013). Gene divergence of homeologous regions associated with a major seed protein content QTL in soybean. *Front. Plant Sci.* 4. doi: 10.3389/fpls.2013.00176

Li, Z., Bachleda, N., Wilson, B., Wood, E. D., Buck, J. W., Carter, T. E., et al. (2022). Registration of G11-7013 soybean germplasm with high meal protein and resistance to soybean cyst nematode, southern root-knot nematode, and stem canker. *J. Plant Regist.* 16, 430–437. doi: 10.1002/plr2.20204

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., and Abe, J. (2008). Genetic redundancy in soybean photoresponses associated with duplication of the phytochrome A gene. *Genetics* 180, 995–1007. doi: 10.1534/genetics.108.092742

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110.20

Mian, M. A. R., Cooper, R. L., and Dorrance, A. E. (2008). Registration of 'Prohio' Soybean. *J. Plant Regist.* 2, 208–210. doi: 10.3198/jpr2007.09.0531crc

Muñoz, F., and Rodriguez, L. S. (2020) *BreedR: Statistical methods for forest genetic resources analysts*. Available at: https://github.com/famuvie/breedR.

Naeve, S., and Miller-Garvin, J. (2021). *United States soybean quality - Annual Report 2021* (St. Paul: University of Minnesota).

Pantalone, V., and Smallwood, C. (2018). Registration of 'TN11-5102' Soybean cultivar with high yield and high protein meal. *J. Plant Regist.* 12, 304–308. doi: 10.3198/jpr2017.10.0074crc

Patil, G., Mian, R., Vuong, T., Pantalone, V., Song, Q., Chen, P., et al. (2017). Molecular mapping and genomics of soybean seed protein: a review and perspective for the future. *Theor. Appl. Genet.* 130, 1975–1991. doi: 10.1007/s00122-017-2955-8

Patil, G., Vuong, T. D., Kale, S., Valliyodan, B., Deshmukh, R., Zhu, C., et al. (2018). Dissecting genomic hotspots underlying seed protein, oil, and sucrose content in an interspecific mapping population of soybean using high-density linkage mapping. *Plant Biotechnol. J.* 16, 1939–1953. doi: 10.1111/pbi.12929

Prenger, E. M., Yates, J., Mian, M. A. R., Buckley, B., Boerma, H. R., and Li, Z. (2019). Introgression of a high protein allele into an elite soybean cultivar results in a high-protein near-isogenic line with yield parity. *Crop Sci.* 59, 2498–2508. doi: 10.2135/cropsci2018.12.0767

Probst, A. H., Laviolette, F. A., Athow, K. L., and Wilcox, J. R. (1971). Registration of protana soybean. *Crop Sci.* 11, 312–312. doi: 10.2135/cropsci1971.0011183x001100020050x

Qi, Z., Pan, J., Han, X., Qi, H., Xin, D., Li, W., et al. (2016). Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. *Mol. Breed.* 36, 1–16. doi: 10.1007/s11032-016-0475-x

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Sebolt, A. M., Shoemaker, R. C., and Diers, B. W. (2000). Analysis of a quantitative trait locus allele from wild soybean that increases seed protein concentration in soybean. *Crop Sci.* 40, 1438–1444. doi: 10.2135/cropsci2000.4051438x

Shannon, G., Chen, P., Crisel, M., Smothers, S., Clubb, M., Vieira, C. C., et al. (2022). S09-13185: High-yield soybean germplasm with elevated protein concentration. *J. Plant Regist.* 16, 417–422. doi: 10.1002/plr2.20169

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of soySNP50K, a high-density genotyping array for soybean. *PloS One* 8, 1–12. doi: 10.1371/journal.pone.0054985

USDA (2023) *Germplasm resources information network (GRIN) - national plant germplasm system*. Available at: https://www.ars-grin.gov/ (Accessed March 7, 2023).

Valliyodan, B., Brown, A. V., Wang, J., Patil, G., Liu, Y., Otyama, P. I., et al. (2021). Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. *Sci. Data* 8, 1–9. doi: 10.1038/s41597-021-00834-w

Vaughn, J. N., Nelson, R. L., Song, Q., Cregan, P. B., and Li, Z. (2014). The genetic architecture of seed composition in soybean is refined by genome-wide association scans across multiple populations. *G3 Genes. Genomes. Genet.* 4, 2283–2294. doi: 10.1534/g3.114.013433

Wang, J., Mao, L., Zeng, Z., Yu, X., Lian, J., Feng, J., et al. (2021). Genetic mapping high protein content QTL from soybean 'Nanxiadou 25' and candidate gene analysis. *BMC Plant Biol.* 21, 1–13. doi: 10.1186/s12870-021-03176-2

Warrington, C. V., Abdel-Haleem, H., Hyten, D. L., Cregan, P. B., Orf, J. H., Killam, A. S., et al. (2015). QTL for seed protein and amino acids in the Benning × Danbaekkong soybean population. *Theor. Appl. Genet.* 128, 839–850. doi: 10.1007/s00122-015-2474-4

Warrington, C., Abdel-Haleem, H., Orf, J. H., Killam, A. S., Bajjalieh, N., Li, Z., et al. (2014). Resource allocation for selection of seed protein and amino acids in soybean. *Crop Sci.* 54, 963–970. doi: 10.2135/cropsci2013.12.0799

Wickham, H. (2016). *ggplot2. Elegant graphics for data analysis* (New York: Springer-Verlag). doi: 10.1002/wics.147

Wilcox, J. R., and Cavins, J. F. (1995). Backcrossing high seed protein to a soybean cultivar. *Crop Sci.* 35, 1036–1041. doi: 10.2135/cropsci1995.0011183X003500040019x

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fpls.2023.1259736

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Zhen Huang,
Northwest A&F University, China
Pritam Kalia,
Indian Agricultural Research Institute (ICAR), India
Jong-In Park,
Sunchon National University, Republic of Korea

*CORRESPONDENCE
Xiaolin Yu
xlyu@zju.edu.cn

†These authors have contributed equally to this work

RECEIVED 16 July 2023
ACCEPTED 13 December 2023
PUBLISHED 08 January 2024

CITATION
Sun N, Chen J, Wang Y, Hussain I, Lei N, Ma X, Li W, Liu K, Yu H, Zhao K, Zhao T, Zhang Y and Yu X (2024) Development and utility of SSR markers based on *Brassica* sp. whole-genome in triangle of U. *Front. Plant Sci.* 14:1259736. doi: 10.3389/fpls.2023.1259736

Development and utility of SSR markers based on *Brassica* sp. whole-genome in triangle of U

Nairan Sun[1,2,3†], Jisuan Chen[4†], Yuqi Wang[1,2,3], Iqbal Hussain[2,3], Na Lei[5], Xinyan Ma[2,3], Weiqiang Li[1,2,3], Kaiwen Liu[2,3], Hongrui Yu[2,3], Kun Zhao[2,3], Tong Zhao[2,3], Yi Zhang[2,3] and Xiaolin Yu[1,2,3*]

[1]Group of Vegetable Breeding, Hainan Institute of Zhejiang University, Sanya, China, [2]Department of Horticulture, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, China, [3]Zhejiang Provincial Key Laboratory of Horticultural Plant Integrative Biology, Hangzhou, China, [4]Department of Supply Chain, Ningbo Haitong Food Technology Co., Ltd., Ningbo, China, [5]Section of Horticulture and Landscape Architecture, Harbin Academy of Agricultural Sciences, Harbin, China

Introduction: Simple sequence repeats (SSR), also known as microsatellites, are crucial molecular markers in both animals and plants. Despite extensive previous research on SSRs, the development of microsatellite markers in Brassica crops remains limited and inefficient.

Methods: Krait software was used to identify microsatellites by genome-wide and marker development based on three recently sequenced basic species of *Brassica* crops in the triangle of U (*Brassica* rapa, *B. nigra* and *B. oleracea*), as well as three allotetraploids (*B. juncea, B. napus* and *B. carinata*) using public databases. Subsequently, the primers and the characteristics of microsatellites for most of them were accordingly designed on each chromosome of each of the six Brassica species, and their physical locations were identified,and the cross-transferability of primers have been carried out. In addition, a B-genome specific SSR marker was screened out.

Results: A total of 79341, 92089, 125443, 173964, 173604, and 222160 SSR loci have been identified from the whole genome sequences of *Brassica* crops within the triangle of U crops, *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. napus* (AACC), *B. juncea* (AABB) and *B. carinata* (BBCC), respectively. Comparing the number distribution of the three allotetraploid SSR loci in the three subgenomes AA, BB and CC, results indicate that the allotetraploid species have significant reduction in the number of SSR loci in the genome compared with their basic diploid counterparts. Moreover, we compared the basic species with their corresponding varieties, and found that the microsatellite characters between the allotetraploids and their corresponding basic species were very similar or almost identical. Subsequently, each of the 40 SSR primers was employed to investigate the polymorphism potential of *B. rapa* (85.27%), *B. nigra* (81.33%) and *B. oleracea* (73.45%), and *B. rapa* was found to have a higher cross-transfer rate among the basic species in the triangle of U. Meanwhile, a B-genome specific SSR marker, *BniSSR23228* possessing the $(AAGGA)_3$ sequence characteristics was obtained, and it located in chromosome B3 with a total length of 97 bp.

Discussion: In this study, results suggest that the pattern of distribution may be highly conserved during the differentiation of basic *Brassica* species and their allotetraploid counterparts. Our data indicated that the allotetraploidization process resulted in a significant reduction in SSR loci in the three subgenomes AA, BB and CC. The reasons may be partial gene dominated chromosomal homologous recombination and rearrangement during the evolution of basic diploid species into allotetraploids. This study provides a basis for future genomics and genetic research on the relatedness of *Brassica* species.

KEYWORDS

Brassica L, simple sequence repeats, microsatellite, primer development, genomewide, B-genome specific SSR marker

1 Introduction

Brassica, as a diverse and important genus within the cruciferous family, includes many important vegetable and oilseed crops for human consumption or food production, such as Chinese cabbage, turnip, cabbage, cauliflower, broccoli, Brussels sprouts, kohlrabi, kale, collards, mustard, and rapeseed. These crops can be stored for a long time and provide sufficient food reserves in winter. Not only several *Brassica* species are economically important oil seeds, spices and vegetables, but also they are rich in essential nutrients such as vitamin C and glucosinolates, which has been associated with a reduced risk of many cancers (Kristal and Lampe, 2002). The genetic relationships between the top six *Brassica* species can be described by the triangle of U model (Nagaharu, 1935) (Figure 1). Therein, three ancestral diploid species *B. rapa* (A genome, n=10), *B. nigra* (B genome, n=8) and *B. oleracea* (C genome. n=9) have been cross-bred over time to produce three allotetraploids: *B. juncea* (AB genome, n=18), *B. napus* (AC genome, n=19) and *B. carinata* (BC genome, n=17). Some *Brassica* crops were also found to be capable in crossing with other important cruciferous crops such as wild radish (*Raphanus*) (Beckie et al., 2003; FitzJohn et al., 2007). This potential to hybridize

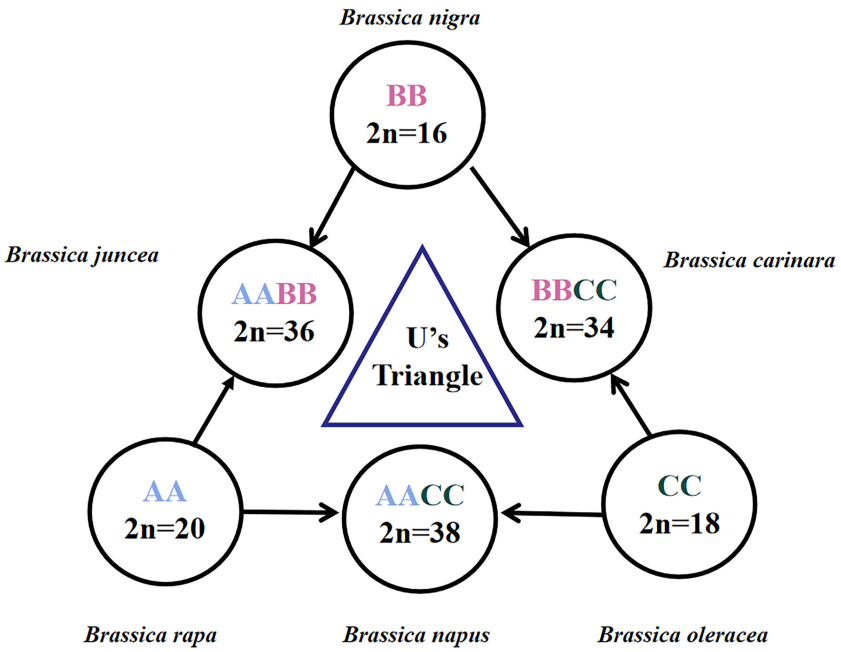


FIGURE 1
Brassica species in the triangle of U. The three diploid basic species are referred to by AA, BB and CC genomes, and the three allotetraploid species are referred to by AABB, AACC and BBCC. The diploid chromosome number (2n) is shown. The image is adapted from U (1935).

with a wide range of inbreds and the diversity of non-domesticated forms of key crop species makes *Brassica* an integral part of global gene banks.

The detection of DNA sequence variation is a crucial step in studying the *Brassica* genome. Over the past two decades, various molecular markers have been used in genetic breeding studies of *Brassica*, such as restriction fragment length polymorphisms (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphisms (AFLP), simple sequence repeats (SSRs), sequence-related amplified polymorphisms (SRAP), sequence-characterized amplified regions (SCAR), and single nucleotide polymorphisms (SNP) (Ananga et al., 2006; Rahman et al., 2010; Zeng et al., 2010; Christensen et al., 2011; Panigrahi et al., 2011; Rezaeizad et al., 2011; Shirasawa et al., 2011). Among these molecular markers, SSRs or microsatellites are characterized by high polymorphism, reproducibility, ease of detection by polymerase chain reaction (PCR), co-dominance, adaptability, transferability, and genomic abundance. Thus, SSRs have been widely used in genetic diversity studies, quantitative trait loci and genetic mapping analysis, gene localization, germplasm classification and evolution and comparative genomics, and it is still one of the important molecular markers in genetic breeding research (Wang et al., 2014).

Traditional methods for developing SSRs involve the probe hybridization of genomic and cDNA libraries containing repetitive motifs, followed by DNA sequencing (Lowe et al., 2004), or the *in silico* analysis of publicly available bacterial artificial chromosome (BAC) sequences (Burgess et al., 2006; Xu et al., 2010), genomic survey sequences and whole-genome shotgun sequences (Cheng et al., 2009; Li et al., 2011). These procedures are time-consuming, costly and labor-intensive; however, with the expansion of DNA sequence information in public databases, the development of SSRs from publicly available DNA sequences has become a rapid and cost-effective alternative (McCouch et al., 2002; Song et al., 2005; Shoemaker et al., 2008). Currently, genome-wide SSR-based development is commonly used in crops such as cocoa, grapes, maize, kidney beans, and prunes (Cai et al., 2009; Cao et al., 2013; Qu and Liu, 2013). This approach has also proved useful in developing SSRs in expressed sequence tags in many agricultural crops, including rice, wheat, cotton, barley, groundnut, cowpea, and radish (Cardle et al., 2000; Kantety et al., 2002; La Rota et al., 2005; Park et al., 2005; Liang et al., 2009; Gupta and Gopalakrishna, 2010; Shirasawa et al., 2011).

With the rapid advancement of whole genome sequencing technology, the genome sequence of cabbage has been released and is available online (http://www.ocri-genomics.org/bolbase/index.html) (Liu et al., 2014). Genome sequences provide a powerful pool of information for genome-wide microsatellite characterization. At the same time, studies on the development of SSRs based on the whole genome of *Brassica* have been limited (Shi et al., 2014). Therefore, in this study, we analyzed the genome-wide SSR information distribution of six *Brassica* species, and located the physical position of SSRs on each chromosome to analyze the relatedness between these species. To evaluate the newly developed genome-wide SSR markers in representative self-crossed lines, we attached these SSR primers of these species as Supplementary Files and screened a number of specific SSR markers by PCR amplification. Moreover, a B-genome specific SSR marker, *BniSSR23228*, was screened out. These results provide great value in relevant research fields including introgression line tracking, genetic diversity analysis, marker-assisted breeding, and so on.

2 Materials and methods

2.1 Source of the whole genome sequence

The genome sequences of three basic species [*B. rapa* (Brara_Chiifu_V3.5), *B. nigra* (Brana_ NI100_V2) and *B. oleracea* (Braol_JZS_V2.0)] and three allotetraploids [*B. juncea* (Braju_tum_V1.5), *B. napus* (Brana_Dar_V5) and *B. carinata* (http://brassicadb.bio2db.com/download.html)] of the genus *Brassica* were downloaded from the *Brassica* Info (http://www.Brassica.info/) website (Chen et al., 2010; Liu et al., 2014). The sequences obtained for *B. rapa, B. nigra, B. oleracea, B. juncea, B. napus* and *B. carinata* were 353140194 bp, 506000232bp, 561157886 bp, 937030072 bp, 850292103 bp, and 1086987601 bp in length, respectively.

2.2 SSR screening

The Krait identification tool was employed to search for the presence of SSR motifs in the genomic sequences (Du et al., 2018). The parameters were set as follows: the minimum number of repeat units was 12 single nucleotides, 7 dinucleotides, 5 trinucleotides, 4 tetranucleotides and pentanucleotides, and 4 hexanucleotides. The frequency and length of the searched SSRs were counted and analyzed.

2.3 Genomic SSR primer design

The primer pairs on both sides of the SSR loci were designed using Krait software (Du et al., 2018). The main parameters were set as follows: the primer length was controlled between 18 and 27 bp, with an optimal size of 20 bp; the melting temperature was 58°C to 65°C, with an optimal temperature of 60°C; the GC content was in the range of 30% to 80%; and the predicted PCR product was in the range of 100-300 bp. All other parameters were set as default.

2.4 Plant material and DNA extraction

The plant materials used in the experiment were from *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. juncea* (AABB), *B. napus* (AACC), *B. carinata* (BBCC), *Raphanus sativus* (RR) and *Arabidopsis thaliana* (At) specimens. The total genomic DNA was extracted from young frozen leaf tissues using the SDS method. The genomic DNA concentrations (ng/μL) were adjusted to an experimentally specific 100 ng/μL using a NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, USA).

2.5 Detection of the transferability of SSR markers

From the SSR primers designed based on the genomes of the six chosen species, 40 pairs of primers were randomly selected in each of the three diploid species, *B. rapa*, *B. nigra* and *B. oleracea*. To perform PCR amplification, a reaction mixture containing 1 μL (100 ng/μL) of template genomic DNA, 1μL (10 μmol/L) of each primer and 12.5 μL of 2×T5 Super PCR Mix (PAGE) buffer was added, followed by the addition of ddH_2O to 25 μL total volume of the reaction mixture. The PCR assay amplification procedure included pre-denaturation at 94°C for 3 min, then 35 cycles involving denaturation at 94°C for 30 s, annealing at 53°C for 30 s, and extension at 72°C for 30 s, and finally, extension at 72°C for 7 min. The PCR reaction procedure was performed on a BIO-RAD S1000TM Thermal Cycler instrument, and samples were stored at 4°C. The SSR cross-transfer rate refers to the number of the amplified bands obtained from the other seven related species except for itself/total bands×100%.

2.6 Polyacryamide gel electrophoresis detection of the PCR product

The PCR product was detected by 12% PAGE following a method modified from "Molecular Cloning: A Laboratory Manual" (Sambrook et al., 2001).

2.7 Statistical analysis

The obtained SSR marker loci were analyzed and calculated using MG2C (http://mg2c.iask.in/mg2c_v2.1/) to locate the physical position of the SSR on each chromosome (Chao et al., 2021). A binary matrix of '1' and '0' was prepared for the SSR marker allele data for all genotypes. The polymorphic information content (PIC) values, gene diversity and heterozygosity were calculated using PowerMarker 3.0 software (Liu and Muse, 2005).

3 Results

3.1 Genome-wide SSR identification of *Brassica* species in the triangle of U

From the genomic sequences of *B. rapa*, *B. nigra*, *B. oleracea*, *B. napus*, *B. juncea*, and *B. carinata* with lengths of 340, 490, 541, 893, 824 and 1085.44 Mb, respectively, we identified 79341, 92089, 125443, 173964, 173604 and 222160 complete mono-nucleotide to hexanucleotide repeat sequence microsatellites with total frequencies of 226, 182, 223.6, 231.31, 235.12 and 204.42 loci per Mb, respectively (Tables 1, 2).

In the genomic SSRs of the six studied species, the distribution of microsatellite motif lengths was almost identical except in *B. nigra* (BB); mononucleotide, dinucleotide, trinucleotide and tetranucleotide repeats accounted for a very similar and relatively

TABLE 1 Distribution of the main SSR types in the genomes of the three basic species.

Motif	*B. rapa*		*B. oleracea*		*B. nigra*	
	Number (%)	Total Length (%)	Number (%)	Total Length (%)	Number (%)	Total Length (%)
Mono	34314 (43.25)	507875 (34.34)	62204 (49.59)	953125 (42.83)	31373 (34.07)	410790 (20.93)
A	32275 (40.68)	470914 (31.84)	57629 (45.94)	862833 (38.78)	31308 (34.00)	409999 (20.89)
C	2039 (2.57)	36961 (2.50)	4575 (3.65)	90292 (4.06)	65 (0.07)	791 (0.04)
Di	28637 (36.09)	663746 (44.88)	41560 (33.13)	892714 (40.12)	38893 (42.23)	1037216 (52.84)
AT	18181 (22.92)	403818 (27.30)	27267 (21.74)	577378 (25.95)	22769 (24.72)	515530 (26.26)
AG	8780 (11.07)	231352 (15.64)	12321 (9.82)	282280 (12.69)	12718 (13.21)	458946 (23.38)
AC	1670 (2.1)	28490 (1.93)	1968 (1.57)	32994 (1.48)	3400 (3.69)	62654 (3.19)
CG	6 (0.01)	86 (0.01)	4 (0.00)	62 (0.00)	6 (0.00)	86 (0.00)
Tri	10667 (13.44)	193242 (13.07)	13652 (10.88)	247191 (11.11)	12929 (14.04)	256068 (13.05)
AAG	3287 (4.14)	58542 (3.96)	4436 (3.54)	80292 (3.60)	4372 (4.75)	86421 (4.40)
AAT	1848 (2.33)	38862 (2.63)	2548 (2.03)	49839 (2.24)	2429 (2.64)	58134 (2.96)
AAC	1398 (1.76)	25053 (1.69)	1338 (1.07)	22425 (1.01)	1453 (1.58)	24603 (1.25)
ACC	720 (0.91)	11982 (0.81)	875 (0.70)	14634 (0.66)	819 (0.89)	13995 (0.71)
ACG	513 (0.65)	8619 (0.58)	520 (0.41)	8652 (0.39)	552 (0.60)	8991 (0.46)
AGG	1106 (1.39)	18759 (1.27)	1604 (1.28)	30024 (1.35)	1210 (1.31)	23403 (1.19)
ATC	1617 (2.04)	28551 (1.93)	2156 (1.72)	38541 (1.73)	1926 (2.09)	37851 (1.93)

(Continued)

TABLE 1 Continued

Motif	*B. rapa*		*B. oleracea*		*B. nigra*	
	Number (%)	Total Length (%)	Number (%)	Total Length (%)	Number (%)	Total Length (%)
CCG	1748 (0.22)	2874 (0.19)	175 (0.14)	2784 (0.13)	168 (0.18)	2670 (0.14)
Tetra	3684 (4.64)	65896 (4.46)	4884 (3.89)	86840 (3.90)	5832 (6.33)	178216 (9.08)
AAAC	390 (0.49)	6680 (0.45)	457 (0.36)	7856 (0.35)	428 (0.46)	7452 (0.38)
AAAG	446 (0.56)	7960 (0.54)	567 (0.45)	10240 (0.46)	564 (0.61)	10472 (0.53)
AAAT	1473 (1.86)	24908 (1.68)	1992 (1.59)	34120 (1.53)	1678 (1.82)	28808 (1.47)
AACT	148 (0.19)	2624 (0.18)	238 (0.19)	4220 (0.19)	147 (0.16)	2468 (0.13)
AATT	167 (0.21)	2800 (0.19)	458 (0.37)	9460 (0.43)	219 (0.24)	3712 (0.19)
ATAC	109 (0.14)	2088 (0.14)	126 (0.10)	2284 (0.10)	131 (0.14)	3648 (0.19)
ATAG	164 (0.21)	4992 (0.34)	304 (0.24)	5476 (0.25)	495 (0.54)	84172 (4.29)
Others	787 (0.99)	13844 (0.94)	742 (0.59)	13184 (0.59)	2170 (2.36)	37484 (1.91)
Penta	1341 (1.69)	28340 (1.92)	1781 (1.42)	37885 (1.70)	1713 (1.86)	37650 (1.92)
AAAAC	119 (0.15)	2515 (0.17)	145 (0.12)	3130 (0.14)	155 (0.17)	3400 (0.17)
AAAAT	234 (0.29)	4940 (0.34)	327 (0.26)	6900 (0.31)	316 (0.34)	6750 (0.34)
AACCG	199 (0.25)	4115 (0.28)	382 (0.30)	7885 (0.35)	157 (0.17)	3245 (0.17)
ACTGG	7 (0.01)	150 (0.01)	17 (0.01)	355 (0.02)	1 (0)	20 (0.00)
Others	782 (0.99)	16620 (1.12)	910 (0.73)	19615 (0.88)	1084 (1.18)	24235 (1.23)
Hexa	698 (0.88)	19932 (1.35)	1362 (1.09)	37452 (1.68)	1349 (1.46)	42990 (2.19)
AAAAAC	45 (0.06)	1440 (0.10)	58 (0.05)	1512 (0.07)	52 (0.06)	1770 (0.09)
AAAAAG	26 (0.03)	690 (0.05)	70 (0.06)	1806 (0.08)	198 (0.22)	5850 (0.30)
AAAAAT	85 (0.11)	2388 (0.16)	86 (0.07)	2244 (0.10)	84 (0.09)	2946 (0.15)
AAAACC	22 (0.03)	576 (0.04)	19 (0.02)	480 (0.02)	14 (0.02)	438 (0.02)
AAAGAG	15 (0.02)	402 (0.03)	20 (0.02)	522 (0.02)	16 (0.02)	402 (0.02)
AAATAT	21 (0.03)	1122 (0.08)	30 (0.02)	822 (0.04)	14 (0.02)	372 (0.02)
Others	484 (0.61)	13314 (0.90)	1079 (0.86)	30066 (1.35)	971 (1.05)	31212 (1.59)
Total	79341 (100)	1479031 (100)	125443 (100)	2225207 (100)	92089 (100)	1962930 (100)

TABLE 2 Distribution of the main SSR types in the genomes of the three allotetraploids.

Motif	*B. juncea*		*B. napus*		*B. carinata*	
	Number (%)	Total Length (%)	Number (%)	Total Length (%)	Number (%)	Total Length (%)
Mono	80518 (46.28)	1177653 (38.33)	79106 (45.57)	1173689 (31.13)	91679 (41.27)	1398206 (25.63)
A	77106 (44.32)	1117561 (36.38)	73785 (42.5)	1061674 (28.16)	89382 (40.23)	1368065 (25.08)
C	3412 (1.96)	60092 (1.96)	5321 (3.07)	112015 (2.97)	2297 (1.03)	30141 (0.55)
Di	56568 (32.52)	1185154 (38.58)	60145 (34.64)	1868030 (49.54)	84030 (37.82)	2815754 (51.32)
AT	29808 (17.13)	567802 (18.48)	37598 (21.66)	946490 (25.10)	54026 (24.32)	1841142 (33.75)
AG	22521 (12.95)	543962 (1.76)	19371 (11.16)	865648 (22.96)	24539 (11.05)	861550 (15.79)
AC	4226 (2.43)	73206 (2.38)	3162 (1.82)	55680 (1.48)	5450 (2.45)	112836 (2.07)
CG	13 (0.01)	184 (0.00)	14 (0.01)	212 (0.01)	15 (0.01)	226 (0.01)

(Continued)

TABLE 2 Continued

Motif	*B. juncea*		*B. napus*		*B. carinata*	
	Number (%)	Total Length (%)	Number (%)	Total Length (%)	Number (%)	Total Length (%)
Tri	23901 (13.74)	435792 (14.18)	22032 (12.69)	460536 (12.21)	26760 (12.05)	632148 (11.59)
AAG	8145 (4.68)	154041 (5.01)	7216 (4.16)	158442 (4.20)	8752 (3.94)	201657 (3.70)
AAT	3661 (2.10)	71178 (2.31)	4110 (2.37)	103179 (2.74)	5038 (2.27)	132789 (2.43)
AAC	2658 (1.53)	45213 (1.47)	2393 (1.38)	42348 (1.12)	2795 (1.26)	51114 (0.94)
ACC	1596 (0.92)	26676 (0.87)	1302 (0.75)	21654 (0.57)	1909 (0.86)	39012 (0.72)
ACG	1098 (0.63)	18114 (0.59)	952 (0.55)	15978 (0.42)	112 (0.05)	1803 (0.03)
AGG	2603 (1.50)	44328 (1.44)	2109 (1.21)	36276 (0.96)	2740 (1.23)	60408 (1.11)
ATC	3822 (2.20)	71109 (2.31)	3631 (2.09)	77541 (2.06)	4091 (1.84)	104343 (1.91)
CCG	318 (0.18)	5133 (0.17)	319 (0.18)	5118 (0.14)	394 (0.18)	23763 (0.44)
Tetra	8570 (4.93)	166296 (5.41)	7791 (4.49)	157596 (4.18)	11798 (5.31)	279644 (5.13)
AAAC	908 (0.52)	15660 (0.51)	833 (0.48)	14284 (0.38)	862 (0.39)	17548 (0.32)
AAAG	1068 (0.61)	19532 (0.64)	927 (0.53)	17536 (0.47)	1157 (0.52)	22680 (0.42)
AAAT	2860 (1.64)	48480 (1.58)	3179 (1.83)	54672 (1.45)	3777 (1.7)	74496 (1.37)
AACT	284 (0.16)	5000 (0.16)	453 (0.22)	8052 (0.21)	75 (0.03)	1756 (0.03)
AATT	337 (0.19)	5812 (0.19)	456 (0.26)	9724 (0.26)	776 (0.35)	22872 (0.42)
ATAC	260 (0.15)	5680 (0.18)	215 (0.12)	4156 (0.11)	259 (0.12)	7848 (0.15)
ATAG	934 (0.54)	33156 (1.08)	469 (0.27)	26228 (0.70)	1570 (0.71)	61660 (1.13)
Others	1919 (1.12)	32976 (1.07)	1259 (0.73)	22944 (0.61)	3322 (1.49)	70784 (1.30)
Penta	2739 (1.57)	59810 (1.95)	2770 (1.60)	60990 (1.62)	4238 (1.91)	134265 (2.46)
AAAAC	295 (0.17)	6420 (0.21)	271 (0.16)	5800 (0.15)	276 (0.12)	6030 (0.11)
AAAAT	518 (0.30)	10915 (0.36)	530 (0.31)	11405 (0.30)	716 (0.32)	19415 (0.36)
AACCG	386 (0.22)	8095 (0.26)	508 (0.29)	10445 (0.28)	671 (0.30)	18425 (0.34)
ACTGG	30 (0.02)	630 (0.02)	49 (0.03)	1005 (0.03)	23 (0.01)	475 (0.01)
Others	1510 (0.86)	33750 (1.10)	1412 (0.84)	32335 (0.86)	2552 (1.16)	89920 (1.65)
Hexa	1668 (0.96)	47610 (1.55)	1760 (1.01)	49902 (1.32)	3655 (1.65)	195090 (3.58)
AAAAAC	134 (0.08)	3960 (0.13)	99 (0.06)	2730 (0.07)	118 (0.05)	3306 (0.06)
AAAAAG	159 (0.09)	4236 (0.14)	80 (0.05)	2034 (0.05)	243 (0.11)	16302 (0.30)
AAAAAT	118 (0.07)	3048 (0.10)	159 (0.09)	4332 (0.11)	327 (0.15)	15234 (0.28)
AAAACC	28 (0.02)	786 (0.03)	23 (0.01)	636 (0.02)	34 (0.02)	2394 (0.04)
AAAGAG	30 (0.02)	822 (0.03)	34 (0.02)	912 (0.02)	44 (0.02)	1200 (0.02)
AAATAT	15 (0.01)	432 (0.01)	36 (0.02)	1626 (0.04)	74 (0.03)	3558 (0.07)
Others	1184 (0.67)	34326 (1.12)	1329 (0.77)	37632 (1.00)	2815 (1.27)	153096 (2.81)
Total	173694 (100)	3072315 (100)	173604 (100)	3770743 (100)	222160 (100)	5455107 (100)

high proportion, while pentanucleotide and hexanucleotide repeats were relatively uncommon. The mononucleotide repeat motifs were the most abundant of the repeat types, with frequencies largely above 40%, and even close to 50% in *B. oleracea* (Figure 2A).

The type distribution of microsatellite motifs was almost identical in the whole genome sequences of *B. rapa*, *B. nigra*, *B. oleracea*, *B. napus*, *B. juncea*, and *B. carinata* (Figure 2B). In other words, the mononucleotide to hexanucleotide motifs making up the major part of the genome sequences of the six *Brassica* species and those that are scarce were essentially the same. From Figure 2B, it is interesting to find that among the mononucleotide repeat sequences, A has the most repetitive motifs; among the

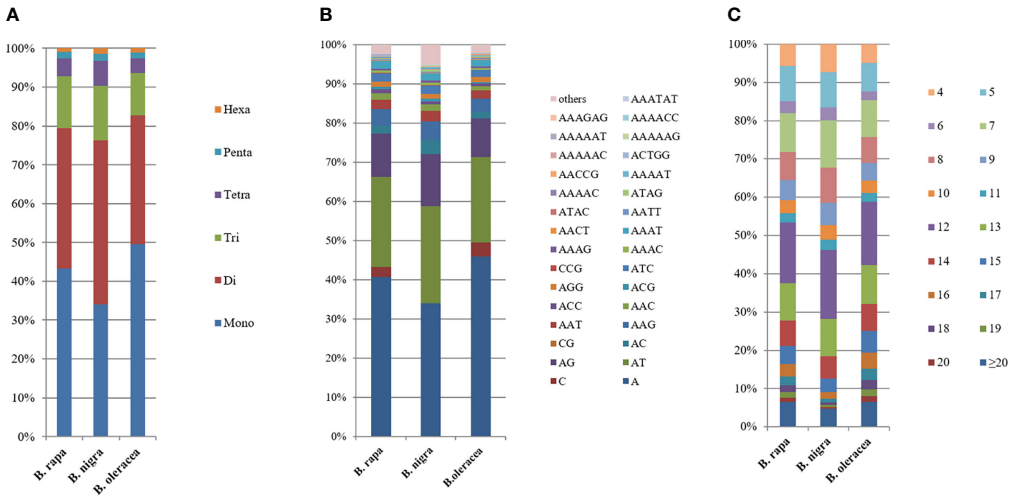


FIGURE 2
Distribution with respect to the motif length **(A)**, type **(B)** and repeat number **(C)** of microsatellites in the whole genomes of *B. rapa*, *B. nigra* and *B. oleracea*. The vertical axis represents the abundance (%) of microsatellites of different motif lengths, types or number of repeats, which are distinguished by different colors. For **(B)**, due to the limited number of items in Excel, the abundance of representative single to pentanucleotide motifs was selected, while the abundance of other motifs was shown in Supplementary Table 2.

dinucleotide repeat sequences, AT has the most repetitive motifs, followed by AG; among the trinucleotide repeat sequences, AAG has the most repetitive sequences, followed by AAC; among the tetranucleotides AAAT has the most repetitive sequences; among the five and six nucleotides, AAAAT and AAAAAT are also more common than other combinations. Most of the single to hexanucleotide sequences that account for the major motifs contained abundant A\T, while the G\C motifs are all among the scarce motifs. This is in good agreement with previous reports of microsatellites identified in *B. rapa, B. oleracea* and *B. napus*. It is also clearly seen that the genomic sequences of *B. rapa* have a much higher content of A\T relative to C\G.

Among the whole genome sequences of *B. rapa* (AA), *B. nigra*, *B. oleracea*, *B. napus*, *B. juncea*, and *B. carinata*, the distribution pattern of the number of motif repeats of microsatellites is essentially the same, except for *B. nigra*, where 12 repeats have the highest proportion of all repeats. (Figure 2C). At the same time, we can see that the microsatellite abundance decreases significantly as the number of motif repeats increases, with the rate of change being the flattest for dinucleotides, followed by single nucleotide as well as trinucleotide repeats, and more drastic changes can be observed for long repeat motifs. (Figures 3, 4).

In addition, we compared the corresponding motif lengths (Figure 5), mono- to hexanucleotide microsatellite numbers and

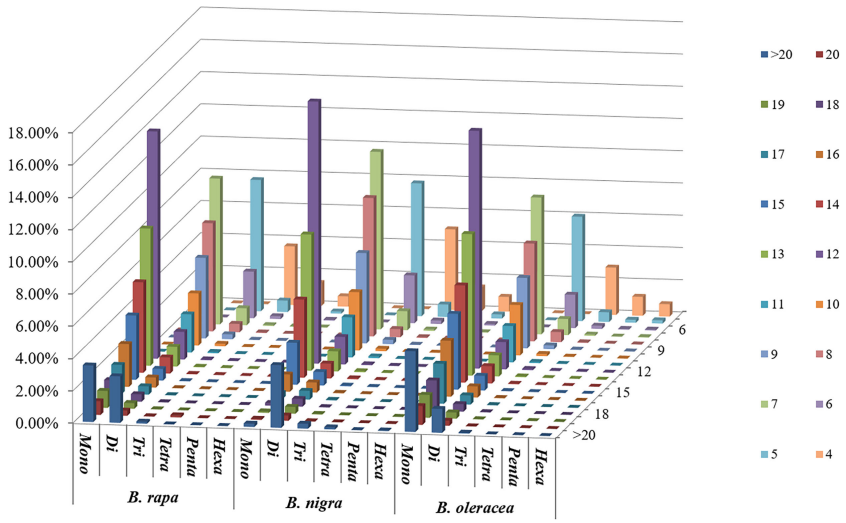


FIGURE 3
Distribution with respect to the motif repeat number of the individual mono- to hexanucleotide repeat microsatellites in the whole genomes of *B. rapa*, *B. nigra* and *B. oleracea*. The vertical axis shows a large number of microsatellites with different motif repeat numbers (from 4 to 20), which are distinguished by different colors.

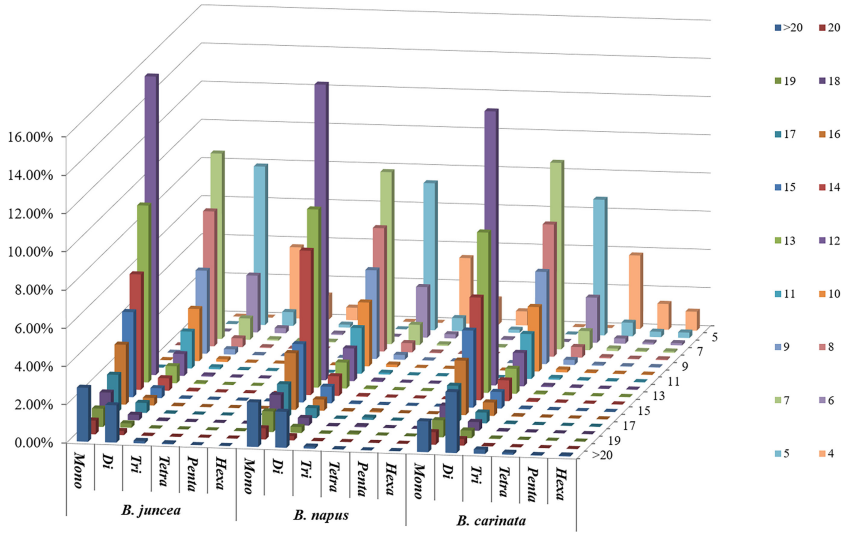


FIGURE 4
Distribution with respect to the motif repeat number of the individual mono- to hexanucleotide repeat microsatellites in the whole genomes of *B. juncea*, *B. napus* and *B. carinata*. The vertical axis shows a large number of microsatellites with different motif repeat numbers (from 4 to 20), which are distinguished by different colors.

motif repeat numbers between the basic species and the two heterotetraploid variants from which they diverged. As can be seen from the figure, the patterns of variation are very similar in the genomic SSRs of the six studied species, but comparisons between them reveal a more similar trend in microsatellite length distribution between *B. rapa* (AA), *B. napus* (AACC) and *B. juncea* (AABB) (Figure 5A); among *B. nigra* (BB), *B. juncea* (AABB) and *B. carinata* (BBCC), the microsatellite length distribution trends are more similar between *B. nigra* (BB) and *B. carinata* (BBCC) (Figure 5B); whereas among *B. oleracea* (CC), *B. napus* (AACC) and *B. carinata* (BBCC), the microsatellite length distribution trends are more similar between *B. oleracea* (CC) and *B. napus* (AACC) (Figure 5C). However, these differences are not highly significant.

By comparing the number distribution of the three allotetraploid SSR loci in the three subgenomes AA, BB and CC, we can find that the allotetraploid species have significant differences in the number of SSR loci in the genome compared with their basic diploid counterparts (Table 3). It is worth mentioning that the reduction in SSR loci is the

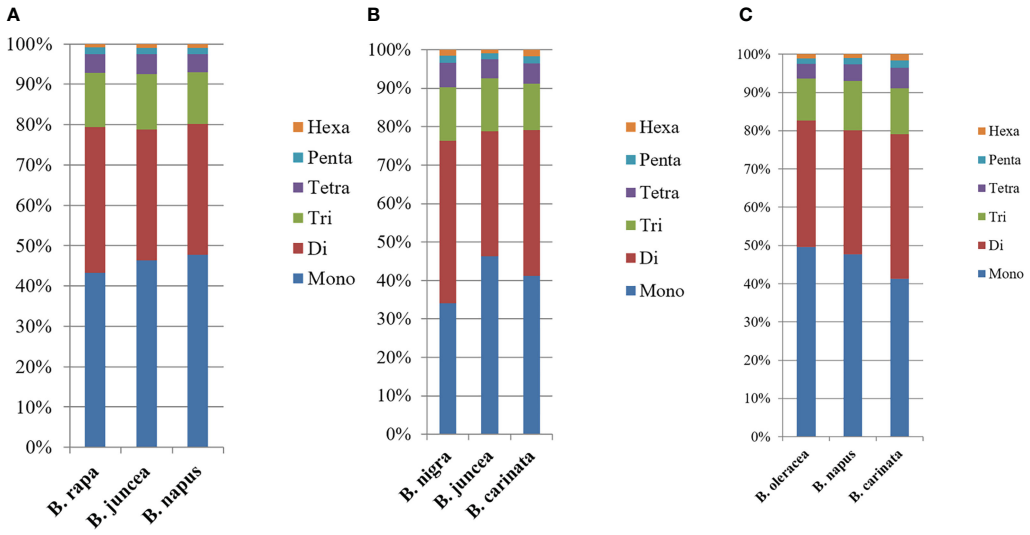


FIGURE 5
Distribution with respect to the motif length of microsatellites in the whole genomes of *Brassica* species in the triangle of U. **(A)** Distribution with respect to the motif length of microsatellites in the whole genomes of *B. rapa*, *B. napus* and *B. juncea*. **(B)** Distribution with respect to the motif length of microsatellites in the whole genomes of *B. nigra*, *B. juncea* and *B. carinata*. **(C)** Distribution with respect to the motif length of microsatellites in the whole genomes of *B. oleracea*, *B. napus* and *B. carinata*. The vertical axis shows the abundance (%) of microsatellites with different motif lengths, which are distinguished by different colors.

10.3389/fpls.2023.1259736

TABLE 3 The number of SSR loci in different genomes in the *B Brassica* species in the triangle of U.

Species	AA genome/ subgenome	BB genome/ subgenome	CC genome/ subgenome
B rapa (AA)	79341	0	0
B nigra (BB)	0	92089	0
B oleracea (CC)	0	0	125443
B juncea (AABB)	73584 (7.3% ↓)	80149 (13.0% ↓)	0
B napus (AACC)	54500 (31.3% ↓)	0	79084 (37.0% ↓)
B carinata (BBCC)	0	90569 (1.7% ↓)	117871 (6.0% ↓)

"↓" means reducing of the number SSR loci.

highest during the allotetraploidization of *B. napus* (AACC), which is 31.3% and 37%, respectively, compared with the diploid AA and CC genome. Furthermore, the reduction in SSR loci is intermediate during the allotetraploidization of *B. juncea* (AABB) L, which is 7.3% and 13.0%, respectively, compared with the diploid AA and BB genome. Moreover, the reduction in SSR loci is the lowest during the allotetraploidization of *B. carinata* (BBCC) L, which is 1.7% and 6.0%, respectively, compared with the diploid BB and CC genome (Table 3). We also compared the distribution of subgenomic chromosome SSR loci in the three basic species and the three allotetraploids respectively, and found that although the number of SSR loci in most chromosomes showed a downward trend, a few chromosomes showed an increase (Supplementary Table 1). For example, in *B. juncea*, the A subgenomic chromosomes A01 and A02 increased by 5.55% and 2.60% compared with the basic species, respectively. Moreover, in *B. carinata*, the B subgenomic chromosomes B01, B03 and B04 increased by 49.53%, 11.17% and 3.13%, respectively (Supplementary Table 1). These results suggest that the cytological and genetic mechanisms of allotetraploid evolution are complex and worthy of further study in the future.

3.2 Distribution and physical location of SSRs on each chromosome of the whole genome of *Brassica* species in the triangle of U

Based on sequencing the whole genome chromosomes of *Brassica* species in the triangle of U, the characteristics of microsatellites on each chromosome of each of the six considered *Brassica* species and their physical location were investigated.

The characteristics of length, type and number of repeats on each chromosome of the six *Brassica* species were consistent with the overall microsatellite characteristics of each species described above. However, the number of microsatellites distributed on each chromosome was extremely heterogeneous. For the basic species *B. rapa* (AA) (Supplementary Table 2), *B. nigra* (BB) (Supplementary Table 3) and *B. oleracea* CC (Supplementary Table 4), the number of microsatellites was the highest on A09 (11214), B02 (13909) and C03 (18286), respectively. In contrast, for the four allotetraploid variants *B. napus* (AACC) (Supplementary Table 5), *B. juncea* (AABB) (Supplementary Table 6) and *B. carinata* (BBCC) (Supplementary Table 7), the microsatellite numbers were the highest on C03 (12561), B02 (12452) and C01 (16036), respectively. This may have occurred because the number of microsatellites is closely related to the length of the chromosomes; the greater the length of a chromosome, the larger the number of its corresponding microsatellites.

In order to explore the exact distribution of microsatellites on each chromosome, the relationship between microsatellites and chromosomes and that between the three basic species and their corresponding allotetraploids should be analyzed more clearly. We used mapping software to locate the physical position of each microsatellite to the corresponding chromosome. The results show that, for the six *Brassica* species, *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. juncea* (AABB), *B. napus* (AACC), and *B. carinata* (BBCC), all chromosomes have higher microsatellite frequencies at and near the ends, and they present lower microsatellite frequencies in and near the middle region. This is consistent with previous studies on the location of microsatellites on chromosomes and may correspond to the distribution around the telomeres and the thylakoids. Secondly, the physical distribution of microsatellites across all chromosomes of the six different species of *Brassica* is highly heterogeneous, suggesting that microsatellites do not occur randomly but their presence is most likely highly correlated with gene function around them. By comparison, we can also find that microsatellite distribution is more concentrated in *B. nigra* (BB) based on the physical position of microsatellites among the three basic species. Meanwhile, in the three allotetraploids, microsatellite distribution is more concentrated in *B. carinata* (BBCC), which may be due to the more concentrated distribution of genes on *B. nigra* (BB) and *B. carinata* (BBCC). The more concentrated distribution of genes on the latter two species is probably related. Moreover, the high concordance between microsatellites and genes strongly suggests the putative role of microsatellites in regulating genome function and in tagging genes using SSR molecules.

3.3 SSR primer design of the whole genome of Br*assica* species in the triangle of U

Using Krait software, primer pairs were successfully designed for each of the six species of *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. napus* (AACC), *B. juncea* (AABB) and *B. carinata* (BBCC), respectively, yielding a total of 52356, 62290, 82984, 111276, 120324 and 144149 primer pairs named in the order of *BrSSR00001 ~ BrSSR52356*, *BniSSR00001 ~ BniSSR62290*, *BolSSR000001 ~ BolSSR082984*, *BnaSSR000001 ~ BnaSSR111276*, *BjuSSR000001 ~*

BjuSSR120324, and *BcaSSR000001* ~ *BcaSSR144149*, respectively. The primer sequence, TM value, SSR motif, expected product length, and start/end position on the chromosome for each SSR marker were determined (Supplementary Tables 8–13).

3.4 Transferability evaluation of the whole genome of B*rassica* species in the triangle of U

In this study, to enrich the SSR marker library of cruciferous crops and confirm the validity of the designed SSR primers, we randomly selected 120 primer pairs (Supplementary Tables 14–16) in three basic species and used genomic DNA from eight cruciferous species as DNA template, namely, *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. juncea* (AABB), *B. napus* (AACC), *B. carinata* (BBCC), *R. sativus* (RR), and *A. thaliana* (At). The cross-transferability of SSR markers from *B. rapa* (40 SSRs), *B. nigra* (40 SSRs) and *B. oleracea* (40 SSRs) was assessed and the affinities of the three basic species in forming three allotetraploids were speculated. We used a total of 120 SSRs, 40 from *B. rapa*, 40 from *B. nigra*, and 40 from *B. oleracea* for cross-amplification studies.

For the 40 primer pairs of *B. rapa*, a total of 310 positive amplifications were made in the eight species, resulting in a total of 325 alleles amplified with a cross-transfer rate of 85.27% (Table 4 and Supplementary Figure 1A). This high cross-transfer rate presumes the validity of these SSR markers for the study of the genomes of other species of cruciferous crops. Among these 325 amplified markers, the PIC values ranged from 0.32 to 0.84 with a mean value of 0.73. The gene diversity ranged from 0.41 to 0.86 with a mean value of 0.7461, and the heterozygosity values ranged from 0.21 to 0.70 with a mean value of 0.55.

For the 40 primer pairs of *B. nigra*, a total of 280 positive amplifications were made across the eight species, resulting in a total of 301 alleles amplified with a cross-transfer rate of 81.33% (Supplementary Table 17 and Supplementary Figure 1B). From this high rate of cross-transfer, we speculate that these SSR markers have some validity for genomic studies of other species of cruciferous crops. The PIC values of these 301 amplified markers ranged from 0.33 to 0.73 with a mean value of 0.50. The gene diversity ranged from 0.38 to 0.71 with a mean value of 0.58, and the heterozygosity values ranged from 0.10 to 0.53 with a mean value of 0.41.

For the 40 primer pairs of *B. oleracea*, a total of 303 positive amplifications were made across the eight species, resulting in a total of 310 amplified alleles with a cross-transfer rate of 73.45% (Supplemntary Table 18 and Supplementary Figure 1C). This indicates that the SSR cross-transfer rate of *B. oleracea* has limitations in terms of its validity for studying the genomes of other species of cruciferous crops. Among these 310 amplified markers, the PIC values ranged from 0.48 to 0.85 with a mean of 0.67. The gene diversity ranged from 0.54 to 0.81 with a mean of 0.64, and the heterozygosity values ranged from 0.0 to 0.71 with a mean of 0.41.

When comparing the polymorphism potential of *B. rapa* (85.27%), *B. nigra* (81.33%) and *B. oleracea* (73.45%), *B. rapa* was found to have a higher cross-transfer rate (85.27%). In addition, the mean values of PIC, genetic diversity and heterozygosity of *B. rapa*-derived SSR markers were relatively high, indicating that among cruciferous plant species, *B. rapa* has better polymorphic potential than *B. nigra* and *B. oleracea*. The cross-species transferability has been demonstrated in *Brassica* crops, while the degree of SSR cross-transfer depends on the evolutionary distance among species (Thakur et al., 2022).

3.5 Application of SSR molecular markers of *Brassica* species in the triangle of U

A B-genome specific SSR marker, *BniSSR23228*, was obtained from 40 selected SSR primers of black mustard (Supplementary Figure 1B). After PCR amplification, polyacrylamide gel electrophoresis, cloning verification screening and sequence alignment were carried out to validate the existence of this specific SSR marker (Figure 6). The validity experiment results indicated that *BniSSR23228* possessing $(AAGGA)_3$ sequence characteristics located in chromosome B3 with a total length of 97 bp (Supplementary Figure 2). Subsequently, this molecular marker can effectively screen the B genome of *Brassica*, and can be used for variety and parent identification, introgression line tracking, genetic diversity analysis, and marker-assisted breeding.

4 Discussion

4.1 Distribution feature of the SSR in the whole genome of *Brassica*

Krait software was used to search for 79341, 92089, 125443, 173964, 173604, and 222160 SSR loci from the whole genome sequences of six species of *Brassica*, namely, *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. napus* (AACC), *B. juncea* (AABB), and *B. carinata* (BBCC), respectively. The frequency of SSR occurrences (average SSRs per Mb) was 226 loci/Mb, 182 loci/Mb, 223.6 loci/Mb, 231.31 loci/Mb, 235.12 loci/Mb and 204.42 loci/Mb, respectively. In a previous report, the PERL5 script MIcroSAtellite (MISA; http://pgrc.ipkgatersleben.de/misa/) was employed for the genomes of *B. rapa*, *B. oleracea* and *B. napus* to obtain 140998, 229389 and 420991 SSR markers, respectively (Shi et al., 2014). Using Karit enabled SSR identification and subsequent primer design in less time than MISA, and the long microsatellites identified in this way were more polymorphic and useful. Among the SSR single nucleotide sequences identified in the six species, *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. napus* (AACC), *B. juncea* (AABB), and *B. carinata* (BBCC), the A sequence repeats were present in 29518 (42.5%), 31308 (34.00%), 57629 (45.94%), 73785 (42.5%), 77106 (44.32%), and 89382 (40.23%) single nucleotides, respectively. They were therefore regarded as the most important of such repeats, while the C sequence was less represented in single nucleotides. This result is consistent with previous studies performing SSR analysis of the whole grapevine genome (Cai et al., 2009). For the dinucleotide repeat type AT, the number of repeats were 13915 (20.04%), 22769 (24.72%), 27267 (21.74%), 37598 (21.66%), 29808 (17.13%) and

10.3389/fpls.2023.1259736

TABLE 4 Amplification results of the *B. rapa* cross-transferability test.

SSR	*B. rapa*	*B. nigra*	*B. oleracea*	*B. juncea*	*B. napus*	*B. carinata*	*A. thaliana*	*R. sativus*
BrSSR00004	+	+	-	+	+	+	+	+
BrSSR01587	+	+	+	+	-	-	+	+
BrSSR03830	+	+	+	+	+	+	+	+
BrSSR03677	+	+	+	+	+	+	+	+
BrSSR01972	+	+	+	+	+	+	+	+
BrSSR30590	+	+	+	+	+	+	+	+
BrSSR06092	+	+	+	+	+	+	-	-
BrSSR09075	+	+	+	+	+	+	-	+
BrSSR14973	+	+	-	+	+	+	+	+
BrSSR15994	+	+	+	+	+	+	-	-
BrSSR13664	+	+	-	+	+	-	-	-
BrSSR11198	+	-	+	+	+	+	+	+
BrSSR17217	+	+	-	+	+	+	+	-
BrSSR18858	+	+	+	+	+	-	-	-
BrSSR16468	+	+	+	+	+	+	+	+
BrSSR18217	+	+	+	+	+	+	+	+
BrSSR18436	+	+	+	+	+	+	+	+
BrSSR21863	+	+	+	+	+	+	+	+
BrSSR23017	+	+	+	+	+	+	+	+
BrSSR24442	+	+	+	+	+	+	+	+
BrSSR28675	+	-	+	+	-	-	-	-
BrSSR25854	+	+	+	+	+	-	-	-
BrSSR31751	+	+	+	+	+	+	-	-
BrSSR34080	+	+	-	+	+	-	+	+
BrSSR33336	+	+	+	+	+	-	+	-
BrSSR34987	+	+	+	+	+	+	+	-
BrSSR36313	+	+	-	+	+	-	-	-
BrSSR37152	+	+	-	+	+	-	-	-
BrSSR37224	+	+	+	+	+	+	+	-
BrSSR36383	+	+	+	+	+	+	+	+
BrSSR38719	+	-	-	-	+	-	-	-
BrSSR39889	+	+	-	+	+	-	+	-
BrSSR42556	+	+	+	+	+	+	+	+
BrSSR44040	+	-	+	+	+	+	-	-
BrSSR44329	+	+	-	+	+	-	-	-
BrSSR45708	+	+	+	+	+	+	-	-
BrSSR47369	+	+	+	+	+	+	+	+
BrSSR48533	+	+	-	+	+	+	+	-

+ means positive result which show the expectant amplification band; - means negative result which show none of the expectant amplification band.

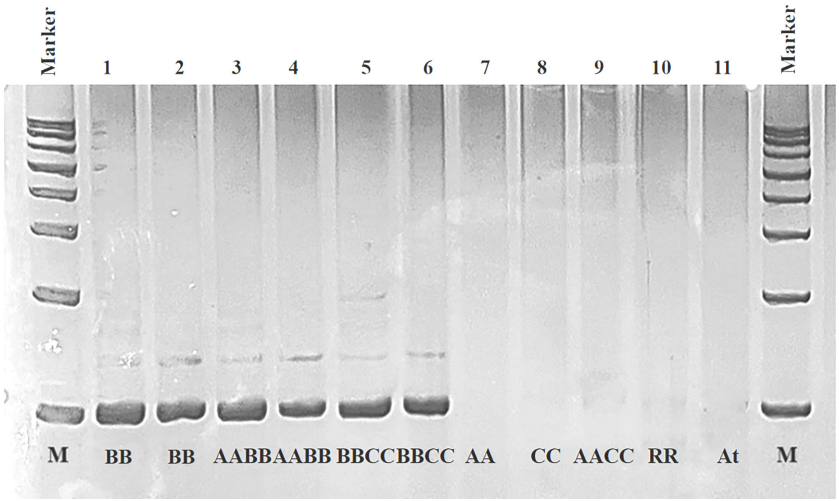


FIGURE 6
PCR amplification result of validity for *BniSSR23228* in different *Brassica* species in the triangle of U. 1, 2: *B. nigra* (BB); 3, 4: *B. juncea* (AABB); 5, 6: *B. carinata* (BBCC); 7, 8, 9, 10, and 11: *B. rapa* (AA), *B. oleracea* (CC), *B. napus* (AACC), *R. sativus* (RR) and *A. thaliana* (At), respectively.

54026 (24.32%) respectively. This is therefore considered as the most significant of this replicate type, consistent with previous studies using SSR analysis of the whole B73 maize genome (Qu and Liu, 2013). The frequency of single nucleotide repeats was the highest of all repeat types, a result that differs from previous SSR analysis studies using MISA in *B. rapa*, *B. oleracea* and *B. napus* (Shi et al., 2014). Moreover, Shi et al. study results showed the microsatellite frequencies of *Brassica*, *Arabidopsis* and other angiosperm species were significantly negatively correlated with both their genome sizes and transposable elements contents (Shi et al., 2013). Qin et al. (2015) investigated the evolutionary regularities of SSRs during the evolution of plant species and the plant kingdom by analysis of twelve sequenced plant genome sequences. The results showed that, SSRs not only had the general pattern in the evolution of plant kingdom, but also were associated with the evolution of the specific genome sequence.

Probably, the deviation may be due to differences in the SSR software algorithms, parameter settings and the original databases used. In addition, our data indicated that the allotetraploidization process resulted in a significant reduction in SSR loci in the three subgenomes AA, BB and CC. The reasons may be partial gene-dominated chromosomal homologous recombination and rearrangement during the evolution of basic diploid species into allotetraploids (Song et al., 2021). Meanwhile, there are a large number of transposable elements (TE) in the *Brassica* species in the triangle of U genome (Cai et al., 2022), and TE insertion seems to result in chromosomal translocation, leading to the reduced number of SSR loci in the three subgenomes AA, BB and CC during the process of allotetraploidization. Of course, further experiments are required to prove this hypothesis. Shi et al. (2014) carried out microsatellite characterization based on genome-wide and marker development in three recently sequenced *Brassica* crops, and suggested that the distribution pattern of microsatellites may be conserved in the genus *Brassica*. This view was reinforced by the use of Krait to identify SSR signatures of six species of *Brassica*, suggesting that the distribution patterns of microsatellites are likely to be conserved in all *Brassica* species. Thus far, the comprehensive identification, characterization and primer development of SSRs for six *Brassica* species in the triangle of U have not been carried out. However, in this study, based on the complete whole genome sequences of six *Brassica* species in the triangle of U, not only were the SSRs of each variety comprehensively analyzed, but also the differences between the *Brassica* crops in the triangle of U were compared and a comprehensive primer design for the SSRs was carried out. To our knowledge, this is the first report to identify the SSR loci and design the SSR primers based on the complete whole genome sequences of six *Brassica* species in the triangle of U together. These markers will act as a powerful tool for future genomic and genetic studies of *Brassica* cruciferous crops in the near future.

4.2 Enrichment of the repertoire of SSR markers of *Brassica* using the cross-transferability approach

High transferability has been reported for SSRs of different plant species, such as “Chiifu” of *B. rapa*, that is, 95% of its SSRs could amplify a fragment of other species (Wang et al., 2011). Thakur et al. (2018) study result indicated 100% cross-transferability was obtained for *B. juncea* and three subspecies of *B. rapa* with 124 *Brassica*-derived SSR loci assayed, while lowest cross-transferability (91.93%) was obtained for *Eruca sativa*. The average % age of cross-transferability across all the seven species was 98.15%. In addition, 47% of EST-SSR markers developed from *B. rapa*, *B. oleracea*, and *B. napus* were transferable to six *Brassica* species (An et al., 2011). Sim et al. randomly selected 41 SSR markers of thistle and alfalfa, and found that the transferability was 53% to 71% in the leguminous plant (alfalfa) and 33% to 44% in the non-leguminous plant (thistle). About 57% of cereal EST-SSRs

could also be amplified in ryegrass (Sim et al., 2009). Additionally, about 60% of EST-SSR markers from barley could be amplified in wheat and rye (Castillo et al., 2008). Cui et al. (2005) used 69 pairs of SSR primers of non-heading Chinese cabbage in eight varieties of *Brassica* crops, and found that the transferability amplification rate was 49.3% to 85.5% and that 33% of the SSR primers in the inter-specific hybrids of *Brassica* presented abundant diversity.

Based on the 1176 SSR-containing ESTs in cabbage, a total of 978 primer pairs have been successfully designed and assessed by validation of the amplification on two inbreed lines (Chen et al., 2010). Subsequently, the results indicated that the developed SSRs from ESTs of *B. oleracea* were valid and practicable in marker-assisted selection and QTL analysis in cabbage (Su et al., 2015). Some useful information about SSR and sequence analysis in *Brassica* crops can also be obtained on the website of *Brassica* DB database (http://Brassica.bbsrc.ac.uk/).

In this work, the functional utility of SSR markers derived from *B. rapa* (AA), *B. nigra* (BB) and *B. oleracea* (CC) was evaluated by analyzing their cross-transferability among *B. rapa* (AA), *B. nigra* (BB), *B. oleracea* (CC), *B. juncea* (AABB), *B. napus* (AACC), *B. carinata* (BBCC), *R. sativus* (RR), and *A. thaliana* (At). From our results, it was inferred that the cross-transferability of SSR markers from *B. rapa* (AA) showed higher potential than those from *B. nigra* (BB) and *B. oleracea* (CC) among these eight species, with cross-transferability rates of 85.27%, 81.33% and 73.45%, respectively. In fact, enriching other varieties with SSR markers alleviates the hassle of the expansion and development process and can facilitate the genetic improvement of new varieties by the genomes of superior varieties. Our findings suggest that genomic SSR markers with high transferability can be used for different *Brassica* species and even non-*Brassica* species. Therefore, these genomic SSR markers with clear location and uniform nomenclature system have high potential to be more widely used in several fields, such as gene localization, genetic mapping, evolutionary analysis, molecular marker-assisted breeding, and provide marker materials for genetic and comparative genomics analysis to further introduce some important agronomic traits into other superior *Brassica* species lacking these traits.

4.3 *BniSSR23228* is a B-genome specific SSR marker

Alien chromosome additions have been used to link species-specific characteristics to particular chromosomes (Kapoor et al., 2011). The plasticity of the *Brassica* genome and existence of natural amphiploids have made it possible to develop several alien chromosome additions by dissecting the *B. rapa*, *B. oleracea*, and *B. nigra* genomes (Chevre et al., 1996; Gu et al., 2009; Li et al., 2013). Compared with the single species-specific SSR marker obtained by comparison between two species developed in previous studies (Gu et al., 2009; Li et al., 2013), the specific SSR developed in this study, *BniSSR23228*, has been verified among *Brassica* species in the triangle of U and their closely related species, radish and *Arabidopsis*. The results revealed that it is a B-genome specific SSR marker, therefore has more significant B-genome specificity and more extensive application value in the future.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://brassicadb.cn.

Author contributions

NS: Methodology, Software, Validation, Writing – original draft, Writing – review & editing. JC: Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing. YW: Formal analysis, Investigation, Methodology, Validation, Writing – review & editing. IH: Data curation, Validation, Writing – review & editing. NL: Funding acquisition, Resources, Validation, Writing – review & editing. XM: Data curation, Investigation, Validation, Writing – review & editing. WL: Data curation, Formal analysis, Investigation, Writing – review & editing. KL: Data curation, Validation, Visualization, Writing – review & editing. HY: Formal analysis, Methodology, Visualization, Writing – review & editing. KZ: Data curation, Investigation, Validation, Visualization, Writing – review & editing. TZ: Data curation, Formal analysis, Investigation, Writing – review & editing. YZ: Data curation, Software, Validation, Writing – review & editing, Writing – original draft. XY: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review & editing, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Project of Sci-tech Foundation of Zhejiang Province (2022C02030 and 2022C02032), the Breeding Project of the Sci-tech Foundation of Zhejiang Province (2021C02065), the Natural Science Foundation of Hainan Province (321MS063), the Project of the Sci-tech Foundation of Ningbo City (2022S189), the Basic Public Welfare Research Plan of Zhejiang Province (LTGN23C150008), Zhejiang Province SanNongJiuFang Science and Technology Cooperation Project (2023SNJF009) and Harbin Academy of Agricultural Sciences & Zhejiang University Agricultural College Research Cooperation Project (2021ZSZZNS03).

Acknowledgments

The authors gratefully acknowledge Prof. Zhenning Liu and Dr. Dongya Wu for stimulating discussion and critical reading of the manuscript. And the authors are grateful to Zhejiang University Press for improving the English.

Conflict of interest

Author JC was employed by the company Ningbo Haitong Food Technology Co., Ltd.

The part of the B-genome specific SSR marker result has been applied for Chinese patent 202310602125.0.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1259736/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Transferability analysis on the designed SSR primers for the three basic species. **(A)**, PCR amplification results of SSR primers for part of the AA genome; **(B)**, PCR amplification results of SSR primers for part of the BB genome; C, PCR amplification results of SSR primers for part of the CC genome.

SUPPLEMENTARY FIGURE 2
Genomic sequence of *BniSSR23228* and sequence alignment results for the BB, AA and CC genomes. **(A)**, Genomic sequence of *BniSSR23228*; **(B)**, Sequence alignment result for the BB genome; **(C)**, Sequence alignment result for the AA genome; **(B)**, Sequence alignment result for the CC genome.

References

An, Z. S., Gao, C. H., Li, J. N., Fu, D. H., Tang, Z. L., and Ortegon, O. (2011). Large-scale development of functional markers in *Brassica* species. *Genome* 54 (9), 763–770. doi: 10.1139/g11-042

Ananga, A. O., Cebert, E., Soliman, K., Kantety, R., Pacumbaba, R. P., and Konan, K. (2006). RAPD markers associated with resistance to blackleg disease in *Brassica* species. *Afr. J. Biotechnol.* 5 (22), 2041–2048. doi: 10.5897/AJB06.594

Beckie, H. J., Warwick, S. I., Nair, H., and Seguin-Swartz, G. S. (2003). Gene flow in commercial fields of herbicide-resistant canola (*Brassica napus*). *Ecol. Appl.* 13 (5), 1276–1294. doi: 10.1890/02-5231

Burgess, B., Mountford, H., Hopkins, C. J., Love, C., Ling, A. E., Spangenberg, G. C., et al. (2006). Identification and characterization of simple sequence repeat (SSR) markers derived in silico from *Brassica oleracea* genome shotgun sequences. *Mol. Ecol. Notes* 6 (4), 1191–1194. doi: 10.1111/j.1471-8286.2006.01488.x

Cai, B., Li, C. H., Yao, Q. H., Zhou, J., Tao, J. M., and Zhang, Z. (2009). Analysis of SSRs in grape genome and development of SSR database. *J. Nanjing Agric. Univ.* 32 (4), 28–32. doi: 10.3321/j.issn:1000-2030.2009.04.006

Cai, X., Lin, R. M., Liang, J. L., King, G. J., Wu, J., and Wang, X. W. (2022). Transposable element insertion: a hidden major source of domesticated phenotypic variation in *Brassica rapa*. *Plant Biotechnol. J.* 20 (7), 1298–1310. doi: 10.1111/pbi.13807

Cao, H. C., Wang, Y., Huang, L. S., Wang, Y. J., Yu, Y. J., and Yang, L. (2013). Large - scale development of SSR markers in the genome of cacao. *J. Shandong Agric. Univ. Nat. Sci.* 44 (3), 340–344. doi: CNKI:SUN:SCHO.0.2013-03-005

Cardle, L., Ramsay, L., Milbourne, D., Macaulay, M., Marshall, D., and Waugh, R. (2000). Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics* 156 (2), 847–854. doi: 10.1093/GENETICS/156.2.847

Castillo, A., Budak, H., Varshney, R. K., Dorado, G., Graner, A., and Hernandez, P. (2008). Transferability and polymorphism of barley EST-SSR markers used for phylogenetic analysis in *Hordeum Chilense*. *BMC Plant Biol.* 8, 9. doi: 10.1186/1471-2229-8-97

Chao, J. T., Li, Z. Y., Sun, Y. H., Aluko, O. O., Wu, X. R., Wang, Q., et al. (2021). MG2C: a user-friendly online tool for drawing genetic maps. *Mol. Hortic.* 1, 16. doi: 10.1186/s43897-021-00020-x

Chen, C., Zhuang, M., Li, K. N., Liu, Y. M., Yang, L. M., Zhang, Y. Y., et al. (2010). Development and utility of EST-SSR marker in cabbage. *Acta Hortic. Sin.* 37 (2), 221–228. doi: 10.16420/j.issn.0513-353x.2010.02.010

Cheng, X. M., Xu, J. S., Xia, S., Gu, J. X., Yang, Y., Fu, J., et al. (2009). Development and genetic mapping of microsatellite markers from genome survey sequences in *Brassica napus*. *Theor. Appl. Genet.* 118 (6), 1121–1131. doi: 10.1007/s00122-009-0967-8

Chevre, A. M., Eber, F., This, P., Barret, P., Tanguy, X., Brun, H., et al. (1996). Characterization of *Brassica nigra* chromosomes and of blackleg resistance in *B-napus-B-nigra* addition lines. *Plant Breed.* 115 (2), 113–118. doi: 10.1111/j.1439-0523.1996.tb00884.x

Christensen, S., von Bothmer, R., Poulsen, G., Maggioni, L., Phillip, M., Andersen, B. A., et al. (2011). AFLP analysis of genetic diversity in leafy kale (*Brassica oleracea* L. convar. *acephala* (DC.) Alef.) landraces, cultivars and wild populations in Europe. *Genet. Resour. Crop Evol.* 58 (5), 657–666. doi: 10.1007/s10722-010-9607-z

Cui, X. M., Hou, X. L., and Dong, Y. X. (2005). Development of SSR primers of non-heading Chineses cabbage and transferability among closely related species. *Sci. Tech. Rev.* 23 (11), 20–23. doi: 10.3321/j.issn:1000-7857.2005.11.006

Du, L. M., Zhang, C., Liu, Q., Zhang, X. Y., and Yue, B. S. (2018). Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics* 34 (4), 681–683. doi: 10.1093/bioinformatics/btx665

FitzJohn, R. G., Armstrong, T. T., Newstrom-Lloyd, L. E., Wilton, A. D., and Cochrane, M. (2007). Hybridisation within *Brassica* and allied genera: evaluation of potential for transgene escape. *Euphytica* 158 (1-2), 209–230. doi: 10.1007/s10681-007-9444-0

Gu, A. X., Wang, Y. H., Xuan, S. X., Chen, X. P., and Shen, S. X. (2009). Establishment of specific SSR from different linkage groups of cabbage compared with Chinese cabbage. *Acta Hortic. Sin.* 36 (8), 1221–1226. doi: 10.16420/j.issn.0513-353x.2009.08.021

Gupta, S. K., and Gopalakrishna, T. (2010). Development of unigene-derived SSR markers in cowpea (*Vigna unguiculata*) and their transferability to other *Vigna* species. *Genome* 53 (7), 508–523. doi: 10.1139/g10-028

Kantety, R. V., La Rota, M., Matthews, D. E., and Sorrells, M. E. (2002). Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol.Biol.* 48 (5), 501–510. doi: 10.1023/a:1014875206165

Kapoor, R., Kaur, G., Banga, S., and Banga, S. S. (2011). Generation of *B. nigra-B. rapa* chromosome addition stocks: cytology and microsatellite markers (SSRs) based characterization. *New Biotech.* 28 (4), 407–417. doi: 10.1016/j.nbt.2010.11.002

Kristal, A. R., and Lampe, J. W. (2002). *Brassica* vegetables and prostate cancer risk: A review of the epidemiological evidence. *Nutr. Cancer* 42 (1), 1–9. doi: 10.1207/s15327914nc421_1

La Rota, M., Kantety, R. V., Yu, J. K., and Sorrells, M. E. (2005). Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. *BMC Genomics* 6, 12. doi: 10.1186/1471-2164-6-23

Li, H. T., Chen, X., Yang, Y., Xu, J. S., Gu, J. X., Fu, J., et al. (2011). Development and genetic mapping of microsatellite markers from whole genome shotgun sequences in *Brassica oleracea*. *Mol. Breed.* 28 (4), 585–596. doi: 10.1007/s11032-010-9509-y

Li, X. J., Wang, Y. H., Xuan, S. X., Zhao, J. J., and Gu, A. X. (2013). Screening of specific SSR markers on different linkage groups of Chinese cabbage compared with cabbage. *J. Plant Genet. Resour.* 14 (4), 694–698. doi: 10.13430/j

Liang, X. Q., Chen, X. P., Hong, Y. B., Liu, H. Y., Zhou, G. Y., Li, S. X., et al. (2009). Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and Arachis wild species. *BMC Plant Biol.* 9, 9. doi: 10.1186/1471-2229-9-35

Liu, K. J., and Muse, S. V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21 (9), 2128–2129. doi: 10.1093/bioinformatics/bti282

Liu, S. Y., Liu, Y. M., Yang, X. H., Tong, C. B., Edwards, D., Parkin, I. A. P., et al. (2014). The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* 5, 11. doi: 10.1038/ncomms4930

Lowe, A. J., Moule, C., Trick, M., and Edwards, K. J. (2004). Efficient large-scale development of microsatellites for marker and mapping applications in *Brassica* crop species. *Theor. Appl. Genet.* 108 (6), 1103–1112. doi: 10.1007/s00122-003-1522-7

McCouch, S. R., Teytelman, L., Xu, Y. B., Lobos, K. B., Clare, K., Walton, M., et al. (2002). Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res.* 9 (6), 199–207. doi: 10.1093/dnares/9.6.199

Nagaharu, U. (1935). Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. *Jpn J. Bot.* 7, 389–452.

Panigrahi, J., Kole, P., and Kole, C. (2011). RFLP mapping of loci controlling self-incompatibility in *Brassica campestris* and their comparative mapping with *B. napus* and *B. oleracea*. *Biol. Plant* 55 (1), 54–60. doi: 10.1007/s10535-011-0007-9

Park, Y. H., Alabady, M. S., Ulloa, M., Sickler, B., Wilkins, T. A., Yu, J., et al. (2005). Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. *Mol. Genet. Genomics* 274 (4), 428–441. doi: 10.1007/s00438-005-0037-0

Qin, Z., Wang, Y., Wang, Q., Li, A., Hou, F., and Zhang, L. (2015). Evolution analysis of simple sequence repeats in plant genome. *PloS One* 10 (12), e0144108. doi: 10.1371/journal.pone.0144108

Qu, J. T., and Liu, J. (2013). A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. *BMC Res. Notes* 6, 403. doi: 10.1186/1756-0500-6-403

Rahman, M., Li, G. Y., Schroeder, D., and McVetty, P. B. E. (2010). Inheritance of seed coat color genes in *Brassica napus* (L.) and tagging the genes using SRAP, SCAR and SNP molecular markers. *Mol. Breed.* 26 (3), 439–453. doi: 10.1007/s11032-009-9384-6

Rezaeizad, A., Wittkop, B., Snowdon, R., Hasan, M., Mohammadi, V., Zali, A., et al. (2011). Identification of QTLs for phenolic compounds in oilseed rape (*Brassica napus* L.) by association mapping using SSR markers. *Euphytica* 177 (3), 335–342. doi: 10.1007/s10681-010-0231-y

Sambrook, J., Russell, D. W., Sambrook, J., and Russell, D. W. (2001). *Molecular cloning: A laboratory manual* (New York: Cold Spring Harbor Laboratory Press).

Shi, J., Huang, S., Fu, D., Yu, J., Wang, X., Hua, W., et al. (2013). Evolutionary dynamics of microsatellite distribution in plants: insight from the comparison of sequenced *Brassica*, *Arabidopsis* and other angiosperm species. *PloS One* 8 (3), e59988. doi: 10.1371/journal.pone.0059988

Shi, J. Q., Huang, S. M., Zhan, J. P., Yu, J. Y., Wang, X. F., Hua, W., et al. (2014). Genome-wide microsatellite characterization and marker development in the sequenced *Brassica* crop species. *DNA Res.* 21 (1), 53–68. doi: 10.1093/dnares/dst040

Shirasawa, K., Oyama, M., Hirakawa, H., Sato, S., Tabata, S., Fujioka, T., et al. (2011). An EST-SSR linkage map of *Raphanus sativus* and comparative genomics of the Brassicaceae. *DNA Res.* 18 (4), 221–232. doi: 10.1093/dnares/dsr013

Shoemaker, R. C., Grant, D., Olson, T., Warren, W. C., Wing, R., Yu, Y., et al. (2008). Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. *Genome* 51 (4), 294–302. doi: 10.1139/g08-010

Sim, S. C., Yu, J. K., Jo, Y. K., Sorrells, M. E., and Jung, G. (2009). Transferability of cereal EST-SSR markers to ryegrass. *Genome* 52 (5), 431–437. doi: 10.1139/g09-019

Song, Q. J., Shi, J. R., Singh, S., Fickus, E. W., Costa, J. M., Lewis, J., et al. (2005). Development and mapping of microsatellite (SSR) markers in wheat. *Theor. Appl. Genet.* 110 (3), 550–560. doi: 10.1007/s00122-004-1871-x

Song, X. M., Wei, Y. P., Xiao, D., Gong, K., Sun, P. C., Ren, Y. M., et al. (2021). *Brassica carinata* genome characterization clarifies U's triangle model of evolution and polyploidy in *Brassica*. *Plant Physiol.* 186 (1), 388–406. doi: 10.1093/plphys/kiab048

Su, Y. B., Liu, Y. M., Li, Z. S., Fang, Z. Y., Yang, L. M., Zhuang, M., et al. (2015). QTL analysis of head splitting resistance in cabbage (*Brassica oleracea* L. var. *capitata*) using SSR and InDel makers based on whole-genome re-sequencing. *PloS One* 10 (9), 17. doi: 10.1371/journal.pone.0138073

Thakur, A. K., Singh, K. H., Sharma, D., Parmar, N., Mishra, D. C., Singh, L., et al. (2022). Enriching the repertoire of SSR markers of *Ethiopian mustard* using cross-transferability approach. *Plant Physiol. Rep.* 27 (1), 65–72. doi: 10.1007/s40502-021-00639-4

Thakur, A. K., Singh, K. H., Singh, L., Nanjundan, J., Khan, Y. J., and Singh, D. (2018). SSR marker variations in *Brassica* species provide in sight into the origin and evolution of *Brassica* amphidiploids. *Hereditas.* 155, 6. doi: 10.1186/s41065-017-0041-5

Wang, G. P., Niu, Y., Wang, W. Y., Yue, S. J., and Lin, J. R. (2014). Transferability of tomato SSR markers to eggplants and other Solanaceous vegetables. *J. South China Agric. Univ.* 35 (4), 56–60. doi: 10.3969/mpb.008.000909

Wang, X. W., Wang, H. Z., Wang, J., Sun, R. F., Wu, J., Liu, S. Y., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nat. Genet.* 43 (10), 1035–U1157. doi: 10.1038/ng.919

Xu, J. S., Qian, X. J., Wang, X. F., Li, R. Y., Cheng, X. M., Yang, Y. A., et al. (2010). Construction of an integrated genetic linkage map for the A genome of *Brassica napus* using SSR markers derived from sequenced BACs in *B. rapa*. *BMC Genomics* 11, 15. doi: 10.1186/1471-2164-11-594

Zeng, X. H., Wen, J., Wan, Z. J., Yi, B., Shen, J. X., Ma, C. Z., et al. (2010). Effects of Bleomycin on microspore embryogenesis in *Brassica napus* and detection of somaclonal variation using AFLP molecular markers. *Plant Cell Tissue Organ Cult.* 101 (1), 23–29. doi: 10.1007/s11240-009-9658-z

frontiers | Frontiers in Plant Science

TYPE Mini Review
PUBLISHED 08 January 2024
DOI 10.3389/fpls.2023.1273859

OPEN ACCESS

EDITED BY
Muhammad Kashif Riaz Khan,
Nuclear Institute for Agriculture and Biology, Pakistan

REVIEWED BY
Sung-Ryul Kim,
International Rice Research Institute (IRRI), Philippines
Sajid Shokat,
International Atomic Energy Agency, Austria

*CORRESPONDENCE
Rubab Zahra Naqvi
rubab.zahra.naqvi@gmail.com
Imran Amin
Imranamin1@yahoo.com
Muhammad Asif
asif.biosafety@gmail.com

†PRESENT ADDRESS
Muhammad Arslan Mahmood,
Division of Plant Sciences, Research School of Biology, The Australian National University, Canberra, ACT, Australia

‡These authors have contributed equally to this work

RECEIVED 07 August 2023
ACCEPTED 08 December 2023
PUBLISHED 08 January 2024

CITATION
Naqvi RZ, Mahmood MA, Mansoor S, Amin I and Asif M (2024) Omics-driven exploration and mining of key functional genes for the improvement of food and fiber crops. *Front. Plant Sci.* 14:1273859. doi: 10.3389/fpls.2023.1273859

Omics-driven exploration and mining of key functional genes for the improvement of food and fiber crops

Rubab Zahra Naqvi[1*‡], Muhammad Arslan Mahmood[1†‡], Shahid Mansoor[1,2], Imran Amin[1*] and Muhammad Asif[1*]

[1]Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering College Pakistan Institute of Engineering and Applied Sciences, Faisalabad, Pakistan, [2]International Center for Chemical and Biological Sciences, University of Karachi, Karachi, Pakistan

The deployment of omics technologies has obtained an incredible boost over the past few decades with the advances in next-generation sequencing (NGS) technologies, innovative bioinformatics tools, and the deluge of available biological information. The major omics technologies in the limelight are genomics, transcriptomics, proteomics, metabolomics, and phenomics. These biotechnological advances have modernized crop breeding and opened new horizons for developing crop varieties with improved traits. The genomes of several crop species are sequenced, and a huge number of genes associated with crucial economic traits have been identified. These identified genes not only provide insights into the understanding of regulatory mechanisms of crop traits but also decipher practical grounds to assist in the molecular breeding of crops. This review discusses the potential of omics technologies for the acquisition of biological information and mining of the genes associated with important agronomic traits in important food and fiber crops, such as wheat, rice, maize, potato, tomato, cassava, and cotton. Different functional genomics approaches for the validation of these important genes are also highlighted. Furthermore, a list of genes discovered by employing omics approaches is being represented as potential targets for genetic modifications by the latest genome engineering methods for the development of climate-resilient crops that would in turn provide great impetus to secure global food security.

KEYWORDS

omics, NGS, crops, agriculture, breeding

Introduction

The majority of efforts to increase crop productivity have focused on conventional breeding techniques, such as phenotyping-based selection. The advancement in genomics over the past 20 years has further boosted the precision and efficiency of breeding programs (Varshney et al., 2005) in many temperate crop species (Eathington et al., 2007). Moreover, the scientific community has invested in the development of genomic resources as well as in intelligent decision support systems (a decision support system that makes extensive use of artificial intelligence (AI)) that result in the reduction of the genotype-phenotype gap and provide effective strategies to develop next-generation climate-resilient crop species (Batley and Edwards, 2016). Being sessile, plants are prone to several stresses that limit their yield. A sound technical knowledge of the gene networks that govern plant stress responses is required to efficiently produce climate-resilient crops. Integrated omics approaches are of great importance as they help in elucidating the essential genetic basis of gene networks that are involved in crop development and plant stress responses (Großkinsky et al., 2018; Muthamilarasan et al., 2019; Naqvi et al., 2022). Omics technologies have been widely utilized to identify the mechanisms involved in plant development, stress responses, yield, and other economically vital traits in important food and fiber crops, such as wheat, rice, maize, potato, tomato, cassava, cotton, etc. In this review, we highlight certain omics-based approaches and their implementation from the perspective of crop improvement. Furthermore, we also described the recent discoveries of crop genomics, transcriptomics, and phenomics and the genes identified through these approaches. Moreover, we have highlighted other technologies (e.g., metabolomics and ionomics) that, if integrated with transcriptomics, can provide deeper insights into the mining of hub genes, which could be employed for developing climate-smart crops. We also provided a list of genes identified from transcriptomics analysis of important food and fiber crops. Lastly, the genes identified from these omics approaches could further be validated through functional genomics techniques, e.g., overexpression, virus-induced gene-silencing (VIGS), and genome editing.

Genomics-assisted breeding for sustainable agriculture

Several types of molecular markers have been employed for crop improvement. Marker-assisted selection using molecular markers greatly increases the speed of crop breeding by allowing traits to be selected without the need to perform phenotyping. The reduced cost, high read accuracy, and long reads of modern sequencing platforms have further enhanced the application of these molecular markers for crop breeding (Kang et al., 2016). For designing tailored crops, one or more of the following genomics-assisted breeding (GAB) approaches, namely marker-assisted recurrent selection (MARS), marker-assisted backcrossing (MABC), advanced backcross quantitative trait loci (AB-QTL), marker-assisted selection (MAS), promotion/removal of allele through genome editing (PAGE/RAGE), haplotype-based breeding, and genomics selection (GS), have been utilized in breeding programs. The initial step for MAS is the identification of specific molecular markers, which are strongly linked with the genomic regions/QTLs regulating the traits of interest. Ultimately, these individual or multiple QTLs can be pyramided through breeding into an elite cultivar through MABC. Successful stories of MABC include the introgression of *QTL-hotspot* into elite varieties of chickpeas for improved yield under drought conditions (Bharadwaj et al., 2021) and improving the yield and stress tolerance in rice variety IR64. This rice variety has improved cooking quality, earliness, high yield, and disease resistance, which has made it registered worldwide (Swamy et al., 2013; Kumar et al., 2014). Other crops such as barley, sorghum, rice, etc. have also been improved for multiple yield and stress-related traits using a similar approach (Hasan et al., 2015; Gorthy et al., 2017; Xu et al., 2018; Cobb et al., 2019; Kim et al., 2021). MAS has also been applied to improve drought tolerance in multiple crops such as maize, rice, sorghum, wheat, sunflower, and soybean (Borrell et al., 2014; Rama Reddy et al., 2014; Khan et al., 2016). Most agronomically valuable genes were cloned by QTL mapping in plants, i.e., by using biparental mapping populations including doubled-haploid libraries (DHLs), recombinant inbred line (RILs), backcross inbred lines (BILs), chromosomal segment substitution lines (CSSLs), fine mapping, and gene validation by using transgenic approaches. Some valuable genes were also cloned by reverse genetics by using insertional mutant pools (Krishnan et al., 2009; Viana et al., 2019).

The GS approach has gained much attention as it enables the selection of traits based on a larger set of markers rather than a few, as in MAS. The examples exhibiting the potential application of GS in cereal breeding included the transfer of eyespot (*Rhizoctonia cerealis*) resistance genes, *Pch1* and *mlo*, for barley powdery mildew, and recessive resistance genes *rym4/rym5* against barley yellow mosaic viruses (Varshney et al., 2021). The evaluation of GS mainly depends on the genomic-estimated breeding values (GEBVs), and to calculate GEBVs, intensive phenotypic and genome-wide marker information is utilized. The benefit of GEBVs is that they allow the prediction of better-performing individuals compared to their parents and are fit for the next breeding cycle; they can also enter directly into the pipeline for variety release (Crossa et al., 2017). The breakthrough success stories in which GS applied for cultivar improvement against diseases include blast in rice, rust in wheat, and bacterial blight (Viana et al., 2019). Moreover, among abiotic stresses, tolerance to salinity, submergence, and drought remained the preferred traits for improvement. Knowledge of specific marker-trait associations is not required for GS. However, the inclusion of a substantial set of markers, such as outcomes of genome-wide association studies (GWAS), into GS models has improved the prediction accuracy (Li et al., 2018). Thus, GS has attracted attention in plant breeding over traditionally employed strategies. With the availability of effective and economical genotyping platforms and advancements in predictive algorithms, GS is anticipated to be a regular method like MAS/MABC in crop breeding programs. The haplotype-based GWAS and selective sweeps are crucial explanations for

understanding genetic diversity in the field of population genetics and genomics, particularly when researching the evolution, adaptability, and stress responses of plant species (Shokat et al., 2020; Bhat et al., 2021; Shokat et al., 2023). A study involving diverse exotics and historical elites developed 2,867 pre-breeding lines for agronomic traits. The study revealed selection footprints and exotic-specific associations, and it uncovered connections specific to invasive species and selection footprints. Many pre-breeding lines contained substantial exotic contributions, despite bias in favor of elite genomes. The selected seven lines were subjected to a varietal release process, and 95 lines have been adopted by national breeding programs for the improvement of the germplasm (Singh et al., 2021). Multiple haplotype and SNP-based model analyses were used to elucidate significant associations within the selection sweeps in tomatoes, which revealed evolutionary insights and potential candidate genes regulating the fruit metabolite content and weight (Zhao et al., 2022). The genomic characterization through NGS and phenotyping data showed 16.1%–25.1% exotic imprints, among which a favorable rare haplotype on chromosome 6D was detected to show minimal grain yield loss upon heat stress. The SNP region annotation showed hits with the isoflavone reductase IRL-like protein of wheat progenitor *Aegilops tauschii*. The overall positive contribution of exotic germplasm was demonstrated, and it was inferred that selected sweeps could be potentially used to secure food insecurity, particularly under climate change threats (Singh et al., 2018).

Pangenomics: capturing the genetic diversity in a species

Increased genomic sequence information from diverse accessions has allowed the development of pangenomes (Zhou et al., 2015; Varshney et al., 2017). Pangenomics is an ideal and comprehensive approach to capturing all the variations in a species as well as representing the combined genetic repertoire of a species. A pangenome generally consists of two components: the core genome and the dispensable genome. Plant studies have discovered that the core genome has a larger size, contributing the maximum portion of genes (Zhao et al., 2018) while the dispensable genome is more likely to contain polymorphic genes, which could account for survival and adaptation in diverse environments. The comparison of the wild species' core genome and the dispensable genome of cultivated species uncovers the effect of domestication (Li et al., 2014). At present, pangenomes of several crops, including wheat, rice, soybean, sesame, and tomato, have been published, revealing structural variations and eliminating the single-sample bias of "reference" genomes. Pangenomics has the capability to exhibit an almost full assessment of the diversification existing in a plant species (Montenegro et al., 2017; Yu J. et al., 2019). Recently, a tomato pangenome has been assembled from 725 phylogenetically and geographically distinct accessions. The recognition of 351 Mbp of sequences that were missing in the reference genome was done using a map-to-pan strategy, which also detected a 4-bp substitution in the *TomLoxC* gene's regulatory region entailing their role in modification in fruit flavor, thus highlighting the selection of fruit quality during the course of domestication (Gao et al., 2019). The advent of robust long-read sequencing technologies and bioinformatics tools is making pangenomics more powerful to aid in discovering crucial genes for trait improvement in major crops.

Exome sequencing applications in crop improvement

Exome sequencing enables researchers to pinpoint important genes involved in the improvement of traits like disease resistance, heat tolerance, and drought resistance by staying focused on the protein-coding portions of the genome. Exome sequencing is utilized for capturing and sequencing 1%–2% of high-value genomic regions, enriched for functional variants and low repetitive regions. It has proven successful in solving biological questions, understanding molecular variation, marker development, and developing genomic resources for complex crop plants (Kaur and Gaikwad, 2017; Bayer et al., 2019; Xiong et al., 2023). Exome-capturing sequencing yielded 27.8 Gb data, identifying 217,948 SNP and 13,554 Indels in wheat, where functionally important SNPs and Indels were identified at 5.0% and 5.3%, respectively. The exome variations in 12 mutant wheat lines provided insights into mutagenic effects, and functionally enriched genes were found in metabolic pathways like plant–pathogen interactions and ADP binding (Li et al., 2022). The G1674A mutation in a barley gene on chromosome 1HL, encoding cellulose synthase-like C1 protein (HvCSLC1), was identified through whole exome sequencing. It was inferred that this mutation leads to the retention of the second intron and premature termination of the HvCSLC1 protein (Gajek et al., 2021). The combined bulk segregant analysis and whole exome-capturing methods employed in potatoes for studying tuber sprout elongation corroborated different QTL sites, helped to narrow down the related genomic regions, and discovered novel QTLs (Sharma et al., 2021). Overall, with this focused strategy, crop development efforts are more precisely made while simultaneously speeding up the breeding process. Exome sequencing, along with other omics technologies, provides breeders with insights that allow them to develop food and fiber crops that can survive in changing environmental conditions, which eventually contributes to a more sustainable and resilient global food supply.

Transcriptomics as a tool to discover vital genes

Transcriptomics aids in investigating the differential gene expression and identification of potential genes involved in response to a particular biotic or abiotic stress. Identification of important genes and elucidation of gene expression is thus a potent strategy to develop crops with improved traits (Abdurakhmonov

et al., 2016). The availability of well-annotated reference genomes through NGS in the postgenomic era has enabled robust transcriptome profiling. RNA-sequencing (RNA-seq) provides a global representation and coverage of differential gene expression, along with the detection of novel transcripts. Several transcriptome studies have shed light on gene and transcript profiling in crop plants. NGS-based transcriptomics has been utilized for all types of RNA with the advances in massively parallel sequencing platforms. NGS-based RNA sequencing techniques include RNA-seq (whole transcriptome quantification or assembly), small RNA-seq (characterization of small RNA, including micro- and noncoding RNA), PRO-seq (detection of nascent RNA), degradome-seq (typically for miRNA target prediction), SMART-seq (quantification of low input RNA), and ScRNA-seq (detection of gene expression in an individual cell) (Dong and Chen, 2013; Olsen and Baryawno, 2018). The latest bioinformatics tools also provide help in the identification of hub genes through weighted co-expression analysis and genome-wide analysis of gene families (Zaidi et al., 2020; Ehsan et al., 2023). Alternative splicing studies through transcriptomics allow the investigation of genetic diversity in different crops (Glushkevich et al., 2022; Farooq et al., 2023). The innovations in NGS technology have empowered gene expression profiling and annotation of transcriptomes in major food and feed crops, including wheat, rice, maize, potato, tomato, cotton, and cassava, under different conditions and stimuli. The important genes identified in recent years by RNA-seq-based transcriptomics linked to certain responses in major crops are highlighted in Table 1.

Metabolomics, ionomics, and proteomics

Metabolites have essential roles in plant growth, development, yield, and defense mechanisms. Metabolite profiling through metabolomics is a vital tool for studying crop interactions with environmental stresses. Different techniques being utilized to study crop metabolites include gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), and nuclear magnetic resonance (NMR), each with their own sample preparation protocols and sensitivity (Pretorius et al., 2021). Metabolomics predicts the biochemical markers linked to phenotypic traits, enabling it to be used as a primary detection tool for the identification of favorable traits, which in combination with genetic analysis can be exploited in crop breeding programs (Peng et al., 2015; Razzaq et al., 2019; Raza, 2020). Comparative metabolomics in the roots and leaves of soybean cultivars (sensitive vs. moderately tolerant) through NMR exhibited primary and secondary metabolites. Among these metabolites, alanine, acetate, citrate, GABA, sucrose, and succinate were found to accumulate in plant roots under flooding conditions, however low levels of these metabolites were detected in leaves (Coutinho et al., 2018). Whitefly-resistant and susceptible cassava accessions were compared through metabolomics, which showed that low levels of lignification are associated with whitefly susceptibility (Perez-Fons et al., 2019).

Ultra-performance liquid chromatography-mass spectrometry (UPLC-MS) has been utilized to study comprehensive metabolite profiling of drought-tolerant and sensitive genotypes of Chinese wheat. Guo et al. showed that seedlings of drought-tolerant wheat genotype harbored higher levels of phenolics and 13-fold higher thymine than drought-sensitive genotype (Guo et al., 2020). GC-MS analysis was done for fatty acids profiling in cottonseed (Illarionova et al., 2020) and NMR-based metabolomics has been used to explore metabolites in Bt vs. non-Bt cotton for insect resistance (Shami et al., 2023).

The advances in functional genomics, along with the availability of statistical and bioinformatics tools, allow metabolic profiling to be used as a phenotypic input for genetic association studies, like QTL, thus facilitating crop improvement. The metabolome analysis of 81 accessions of barley under drought and heat stress revealed 57 metabolite QTLs, which were mostly involved in antioxidant defense responses (Templer et al., 2017). Metabolite-based GWAS is another powerful tool to link genetic factors with primary and secondary metabolites. It provides a prospect for identifying candidate genes by exploiting the information from integrated genetics and metabolites. This approach was used efficaciously in tomatoes and detected 44 loci associated with fruit metabolites (Sauvage et al., 2014). mGWAS in 175 rice accessions showed 323 associations among SNPs and metabolites (Matsuda et al., 2015). Another mGWAS study displayed 16 metabolites related to threonine-producing genes in rice under abiotic stress (Muthuramalingam et al., 2018). Thus, metabolomics has great potential to identify candidate genes and quantitative loci that can be used for crop improvement.

Ionomics is another powerful approach, introduced around a decade ago which provides information on the metabolism of elemental composition in plants. It is a high-throughput technique to study the organism's molecular mechanistic basis of mineral nutrients and their trace element components (also termed the ionome) (Huang and Salt, 2016). For instance, the functional analysis of wheat ionome showed variation in sulfur and phosphorous content associated with grain's phenotype (Fatiukha et al., 2020). Furthermore, the genome-ionome linkage study in rice revealed 12 micronutrients linked to brown rice, which exhibited its nutrient-dense properties (Pasion et al., 2023). Ionome study combined with GWAS and QTL analysis has shown that shoot and root ionomes in rice were associated with 114 genomic regions where the most significant regions were associated with cadmium, manganese, molybdenum, and sulfur, thus displaying the strength of this approach to manipulate and interrogate the complex traits (Cobb et al., 2021). Ionome and transcriptome combined analysis of two cotton varieties under salinity stress showed accumulation variation of different nutrients in different plant tissues and expressional changes in ion transport-related genes (Guo H. et al., 2019).

Proteomics allows for the study of expressed proteins in crops under specific conditions. A combination of crop proteomics with

TABLE 1 Potential gene targets identified via transcriptomics in food and fiber crops.

Variety	Condition or stress	Tissue	Sequencing platform	Approach	No. of DEGs or variants	Important genes/pathways	Reference
Wheat							
Nongda 015 and FZ30	Powdery mildew	Leaf	Illumina HiSeq 4000	2-step bulked segregant RNA sequencing (BSR-Seq)	31 and 20	*Pm5e*	(Xie et al., 2020)
Yunong211	Dithiothreitol and tauroursodeoxycholic acid for endoplasmic reticulum stress	Seedling	Illumina HiSeq	RNA-Seq	8,204	Photosynthesis-related genes, antioxidants, phytohormones, transcription factors	(Yu X. et al., 2019)
PBW677 and PBW703	Nitrogen use efficiency	Root and shoot	Illumina Nextseq500	RNA-Seq	2,406	ABC and SWEET transporters, MYB, bHLH, WRKY, zinc-finger nuclease	(Kaur et al., 2022)
Zhengmai 366 and Chuanmai 42	Drought	Root	Illumina HiSeq 6000	RNA-Seq	11,083	16 dehydrin genes	(Xi et al., 2023)
Rice							
IR36 and Weigu	Salinity	Bursting bud	Illumina HiSeq X Ten	RNA-Seq and QTL-Seq	5	*OsSAP16*	(Lei et al., 2020)
Sahabhagidhan and Geetanjali	Cold	Leaf	Illumina HiSeq2000	RNA-Seq	13,930 and 10,599	AP2/ERF, MYB, WRKY	(Pradhan et al., 2019)
02428 and YZX	Seed vigor	Seed and seedling	Illumina HiSeq	GWAS, QTL, and RNA-Seq	44	*OsEXPA17*, *OsLEA4*, *hsp20*, *OsGH3.8*, GA, and IAA-responsive genes	(Guo T. et al., 2019)
IR64 and Apo	Drought	Leaf	Illumina GAIIx	RNA-Seq	170 and 4	Dehydrin, MYB, NAC, zinc finger, bZIP, HSF-type DNA-binding protein	(Ereful et al., 2020)
Maize							
Zao 8-3 and Ji 853	Low temperature	Seed embryo	Illumina NovaSeq 6000	GWAS and RNA-Seq	10	MAPK and fatty acid metabolism	(Zhang et al., 2020)
B73	Nitrogen stress	Seedling	Illumina HiSeq 2500	Small RNA-Seq	226	miR169, miR398, miR408, miR1214, miR2199	(Yang et al., 2019)
K12 and W64A	Deep seeding	Mesocotyl	Illumina NovaSeq	BSA-Seq and RNA-Seq	24	Cell wall, phytohormones, circadian clock-related genes	(Zhao and Niu, 2022)
Potato							
Kufri Gaurav	Nitrogen use efficiency	Leaf, root, and stolon	Ion Proton	RNA-Seq	206, 144, 775	Superoxide dismutase, GDSL esterase lipase, proline-rich proteins, probable phosphatase 2C, nitrate and sugar transporters, SPX domain, VQ motif, bHLH	(Tiwari et al., 2020)
Longshu No. 3	Wound healing	Tuber	Illumina HiSeq 2500	RNA-Seq	7,665	WRKY, NAC, MYB, sugar and starch metabolism, phytohormone regulation	(Jiang et al., 2022)
Vanderplank and Innovator	Powdery scab	Tuber	Illumina HiSeq 2000	RNA-Seq	2,058	StMRNA, StWRKY6, StUDP, StLOX, StSN1, StPRF	(Lekota et al., 2019)
Tomato							
LA1698 and LA2093	Heat	Leaf	BGISEQ-500	RNA-Seq and QTL-Seq	23,458	SlCathB2, SlGST, SlUBC5, and SlARG1	(Wen et al., 2019)

(Continued)

TABLE 1 Continued

Variety	Condition or stress	Tissue	Sequencing platform	Approach	No. of DEGs or variants	Important genes/pathways	Reference
Moneymaker	Short- and long-term hypoxia	Root	Illumina Nextseq500	RNA-Seq	267 and 1,421	CS9, RBOHB, CAT, MT2B, and ACO1	(Safavi-Rizi et al., 2020)
Local variety	Heat	Leaf	Illumina HiSeq 2500	RNA-Seq and proteome analysis	91	HSPs, HSFs, BAGs, NAC, MBF1C	(Ding et al., 2020)
Ailsa Craig and SlBES1-RNAi-8	Fruit softening	Fruit	Illumina Miseq	RNA-Seq	24	SlBES1 and PMEU1	(Liu et al., 2021)
Cassava							
South China 6068	Waterlogging	Leaf and Root	Illumina	RNA-Seq	2,538 and 13,364	MYBs, WRKYs, NACs, AP2/ERFs, glycolysis, photosynthesis, and galactose metabolism	(Cao et al., 2022)
8 cassava varieties	Cassava brown streak disease	Leaf	Illumina HiSeq 2500	RNA-Seq	8,971	Cinnamic acid, PAL1, PAL2, and chalcone synthase	(Kavil et al., 2021)
Arg7 and W14	Abiotic and biotic stresses	Leaf, stem, and root	Illumina GA II	RNA-Seq	91	MePOD genes	(Wu et al., 2019)
Cotton							
G. hirsutum acc. TM-1 and *G. barbadense* cv. Hai7124 and acc. 3-79	Fiber development	Buds	Illumina Novaseq	RNA-Seq and co-expression analysis	1,850 and 1,050	GhP2C72, bHLH, MYB, GhIAA16, HD-ZIP, TCP, GhARF2b, WRKY	(Zhang J. et al., 2022)
G. arboreum (Ravi)	Whitefly-mediated CLCuD	Leaf	Illumina HiSeq 2500	RNA-Seq and co-expression analysis	50	CRT, β-1,3-glucanase, HSP40, HSP70, NADH, COX1, COX3, MYB, NRT1/PTR family	(Naqvi et al., 2017)
G. arboreum (FDH 228)	Drought and whitefly	Leaf	PacBio IsoSeq and Illumina	RNA-Seq	1,343	CRT1, ERF, bZIP, bHLH, ColI, JAZ1, WRKY, MAPK	(Farooq et al., 2023)
G. hirsutum (Karishma)	Whitefly-mediated CLCuD	Leaf	Illumina HiSeq 2500	RNA-Seq and co-expression analysis	53	AOS, MYB, NAC, bHLH, Auxin, cytokinin, ABA, ethyltransferases	(Naqvi et al., 2019)
G. hirsutum (Mac7)	Whitefly-mediated CLCuD	Leaf	Illumina HiSeq 2500	RNA-Seq and co-expression analysis	55	NRT1/PTR family, nitrate reductase, IAA4, SAUR-36, cytochrome P450, E3 ubiquitin-protein ligase	(Zaidi et al., 2020; Aslam et al., 2022)
G. hirsutum SG747 and *G. barbadense* Giza75	Oil accumulation	Ovule	Illumina HiSeq 2500	RNA-Seq and co-expression analysis	14	*GhCYSD1*, *TAG*, *FAD3*, *BGAL*	(Song et al., 2022)

advanced phenomics and other omics technologies can further assist in the breeding of climate-smart crops (Komatsu et al., 2013). Proteomic studies most predominantly use two-dimensional gel electrophoresis (2-DE) and liquid chromatography (LC)-based techniques that bring forth the proteomes as well as post-translational modifications. Proteomic analysis of soybean varieties by a 2-DE-based procedure under drought and heat stress demonstrated 25 important proteins (Das et al., 2016). The combined metabolome and proteome of maize inbred lines and hybrids showed an abundance of photosynthesis-related proteins, depicting the correlation of hybrid vigor with efficient removal of toxic compounds in hybrids through photorespiration and higher levels of photosynthesis (Li et al., 2020). Comparative proteomics in two rice cultivars under H_2O_2 stress revealed proteins related to oxidative metabolism, photosynthesis, and cell defense mechanisms (Bhattacharjee et al., 2023). Metabolomics coupled with proteomics in cassava cultivars under Sri Lankan cassava mosaic virus stress linked results from both approaches and identified pathways involved in plant viral interactions (Siriwan et al., 2023). Thus, proteomics can deliver candidate genes that could be utilized for marker-assisted breeding programs (Jan et al., 2023).

Phenomics, artificial intelligence, and speed breeding

Phenotypic information is crucial to be utilized in crop breeding; however, recording the phenotypic information in breeding programs remains laborious and time-consuming. The advances in high-throughput computing, remote sensing, artificial intelligence, machine learning, and robotics have made automated phenotyping possible through an approach known as phenomics (Ohyanagi et al., 2022). High-throughput phenomics allows for the measurement of different plant traits, including stress and disease, with automation and precision. A phenomics-based collection of large datasets can be handled, analyzed, and interpreted by modern machine learning algorithms to gain useful intuitions and future predictions of incidence. Neural networks, vector machines, and k-nearest neighbors have been employed in maize, soybean, and wheat for the detection and classification of insect pests (Kasinathan et al., 2021). Hyperspectral imaging, nonimaging spectroscopy, and red–green–blue (RGB) imaging based automated techniques have been emphasized as potential methods for real-time differentiation between crops and weeds in the field for timely management of the weeds (Su, 2020). Artificial neural network-based classification was used to detect blast disease in rice plants with 100% accuracy (Ramesh and Vydeki, 2019). Unmanned aerial vehicle (UAV) imaging and support vector machine classification were used for the crop's texture information for crop monitoring and yield forecasting (Kwak and Park, 2019). Paudel et al. exploited machine learning models on five crops, including barley, potato, sunflower, soft wheat, and sugar beet in the Netherlands, Germany, and France, which provide workflows to forecast crop yields (Paudel et al., 2021). Hitech phenomics is also aiding in identifying nutrient deficiency and water scarcity in crop-cultivated lands (Sahoo et al., 2023). Another innovation of recent years, speed breeding, i.e., attaining multiple crop generations with reduced generation time under controlled conditions, is an influential approach for efficient plant breeding. Speed breeding, along with advanced AI, provides a platform to accelerate plant breeding programs via linking phenomics and genomics, particularly under climate-changing scenario (Rai, 2022). Recent innovations in precision agricultural technologies like remote sensing, the Internet of Things (IoT), and machine learning can help breeders and farmers make informed decisions and optimize their farming practices. These advanced technologies can play a significant role in sustainable agriculture by improving crop yield, reducing resource wastage, and enhancing overall efficiency (Naqvi et al., 2020).

Functional genomics approaches for tailored crop improvement

Most of the agronomically important traits are of complex inheritance and challenging to improve. In this case, the mutant and variant alleles can be identified by wide-association studies and QTL mapping (as discussed above), which further need to be functionally validated before being utilized in the breeding program. Understanding the molecular, genetic, and functional basis of a particular gene can help breeders and researchers develop climate-resilient, more productive, and stress-tolerant cultivars.

Conventionally, mutagenesis is an important strategy to introduce mutations, which can be used as a tool for gene functional study and to develop genetic variability. Moreover, to evaluate the mutants and understand gene function, either a forward genetics (from phenotype to genotype) or reverse genetics (genotype to phenotype) strategy can be utilized. Eliminating gene expression or disrupting gene structure exhibits morphological changes in phenotype, providing evidence of the relationship between a gene and its biological function. Although the spontaneous mutation rates are very low (approximately 10^{-5} to 10^{-8}) in plants, but mutagenesis is not always effective in gene functional analysis of (Varshney et al., 2005) those genes that are only required under specific biotic and abiotic stress; (2) those genes which are involved in growth and development; and (3) redundant genes because losing these gene function may not lead to morphological changes (Jiang and Ramachandran, 2010; Wang et al., 2013; Viana et al., 2019).

Another strategy of functional genomics is insertional mutagenesis, which includes transfer DNA (T-DNA) insertions, retrotransposon, and transposon tagging. These strategies have been widely used in developing rice mutant libraries. In an analysis of 206,668 insertion flanking sequence tags (FSTs), it was found that 32,459 rice genes have already had insertion tags, and about 50% of predicted protein-coding genes have been equipped with insertional mutagenesis. This study showed the importance of insertional mutagenesis but also had some drawbacks, such as manual manipulations and high cost. However, new tools with more directed, gene-specific methods are needed.

Over the past decade, several genes with substantial phenotypic effects have been functionally validated in different crops via clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (CRISPR/Cas9)-based genome editing (GE) to improve crop performance against changing climatic conditions. A key feature of CRISPR/Cas9 is the generation of double-stranded breaks (DSBs) of DNA at target loci, which can further be repaired by two cellular mechanisms: nonhomologous end joining (NHEJ) and homology-directed repair (HDR). This tool offers to target various sites simultaneously by utilizing multiple sgRNAs while expressing a single Cas9 or Cpf1 protein (Chen et al., 2019).

Crop-specific functional genes have been exploited to generate gene-edited crops, and approximately more than 60 success stories have been published for drought tolerance, better cell-wall expansion, improved oil quality, and other plant traits. Furthermore, the crop genes that have been exploited by the pathogens for virulence and pathogenicity can be targeted through CRISPR/Cas9, providing an opportunity to break the life cycle of the pathogen (Mahmood et al., 2022).

Several CRISPR-Cas nucleases and their engineered variants have been momentously expanded beyond generating double-stranded DNA breaks (Huang and Puchta, 2021). This technology has advanced immensely owing to *Cas* variants and gene editing

10.3389/fpls.2023.1273859

approaches aided by apt bioinformatics pipelines. For instance, Cas9 and Cas12a systems have recognized different protospacer-adjacent motif (PAM) for the diagnosis of DNA and RNA viruses (Zhu et al., 2020), while the SHERLOCK system has been employed in soybeans for genotyping and quantification of different traits using crude extracts (Abudayyeh et al., 2019).

Through genomics and transcriptomics data, it has now become possible to screen vital genes systematically. This is possible by using silencing tools such as RNA interference (RNAi) and VIGS, which reduce the expression of specific host target genes and accelerate the plant's functional genomics. As recently reviewed (Lacomme, 2015; Mahmood et al., 2023), many agricultural VIGS vectors derived from both DNA and RNA viruses are presently available for a wide range of plant species to knock out/down gene expression for functional genomics. The innovative virus-induced genome editing (VIGE) approach is an upgrade of VIGS based on a CRISPR system that offers gene editing with higher efficiency without typical laborious transformation protocols (Zhang C. et al., 2022).

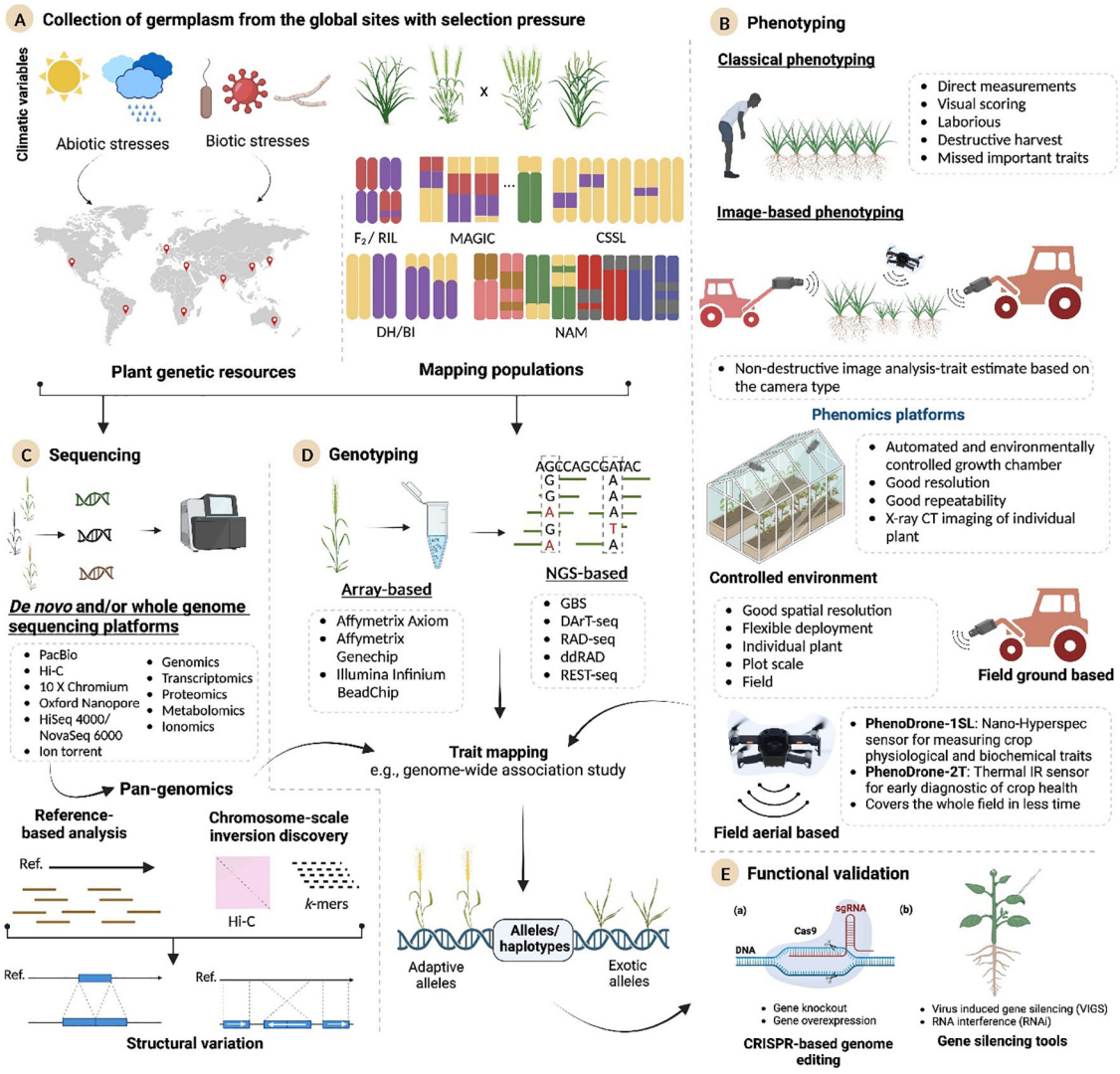


FIGURE 1
Integrated omics technologies for food and fiber crop improvement. The schematic exhibits a holistic approach that aims to identify the favorable alleles or purge the deleterious alleles in the plant genome for designing climate-resilient crops. **(A)** The plant genetic resources and experimental populations from sites experiencing natural selection pressure are selected to serve as valuable sources for genetic variations. **(B)** Phenotyping approaches consist of classical and high-throughput methods. The advanced imaging platforms span from those operating under controlled conditions to field-based conditions. **(C)** Long-read sequencing methods provide high-quality reference genomes and facilitate pangenomic analysis. **(D)** Advanced high-throughput genotyping approaches develop genome-wide marker information on these panels. **(E)** New genes/haplotypes discovered from analyzing sequence information will be further validated by using the VIGS or CRISPR-Cas system, paving the way forward for enhanced food and fiber crop improvement. MAGIC, multiparent advanced generation intercross; NAM, nested assisted mapping; DHLs, doubled-haploid libraries; RILs, recombinant inbred line; BILs, backcross inbred lines; CSSLs, chromosomal segment substitution lines; VIGS, virus-induced gene silencing; VIGE, virus-induced gene editing; GBS, genotype-by-sequencing; RAD-seq, restriction site-associated DNA sequencing; REST-seq, restriction fragment sequencing.

Conclusions and future prospects

State-of-the-art sequencing and bioinformatics approaches are being widely used to explore genetic variations in crops. These advances have paved the way for the exploitation of omics technologies such as genomics, pangenomics, transcriptomics, metabolomics, ionomics, proteomics, and phenomics for the identification of potential molecular markers and genes for crop improvement. Functional validation of these genes is possible using VIGS or GE approaches. Identification of genes/markers using integrated omics technologies has the potential to greatly enhance trait selection and, when combined with speed breeding, significantly accelerate crop improvement. In the era of food insecurity and climate change, interconnected utilization of omics technologies, artificial intelligence, speed breeding, and genome editing (Figure 1) can certainly revolutionize breeding programs to produce climate-smart food and fiber crops for meeting zero hunger and feeding millions of people across the globe. The unprecedented ability of CRISPR/Cas9 technology has led to the tremendous advances in basic plant research and crop improvement. Certain prospects, such as (1) CRISPR/Cas-mediated multiplex gene regulation as a potential plant synthetic biology tool; (2) exploring cop wild relatives (CWRs) by employing omics technology; (3) improved CRISPR/Cas delivery systems; (4) improved gene editing efficiency by HDR mechanism; and (5) GMO regulatory landscape and concerns, have still been the bottlenecks in the development of climate-resilient and future-smart crops.

Author contributions

RN: Conceptualization, Data curation, Project administration, Writing – original draft, Writing – review & editing. MM: Data curation, Writing – original draft, Writing – review & editing. SM: Writing – review & editing. IA: Conceptualization, Project administration, Supervision, Writing – review & editing. MA: Conceptualization, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We thank Dr. Julian R. Greenwood from the Plant Sciences Division at the Australian National University, Australia, for constructive feedback, valuable suggestions, and insightful comments that helped us improve this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abdurakhmonov, I. Y., Ayubov, M. S., Ubaydullaeva, K. A., Buriev, Z. T., Shermatov, S. E., Ruziboev, H. S., et al. (2016). RNA interference for functional genomics and improvement of cotton (Gossypium sp.). *Front. Plant Sci.* 7, 202. doi: 10.3389/fpls.2016.00202

Abudayyeh, O. O., Gootenberg, J. S., Franklin, B., Koob, J., Kellner, M. J., Ladha, A., et al. (2019). A cytosine deaminase for programmable single-base RNA editing. *Science* 365 (6451), 382–386. doi: 10.1126/science.aax7063

Aslam, M. Q., Naqvi, R. Z., Zaidi, S.-e., Asif, M., Akhter, K. P., Scheffler, B. E., et al. (2022). Analysis of a tetraploid cotton line Mac7 transcriptome reveals mechanisms underlying resistance against the whitefly Bemisia tabaci. *Gene* 820, 146200. doi: 10.1016/j.gene.2022.146200

Batley, J., and Edwards, D. (2016). The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Curr. Opin. Plant Biol.* 30, 78–81. doi: 10.1016/j.pbi.2016.02.002

Bayer, M., Morris, J. A., Booth, C., Booth, A., Uzrek, N., Russell, J. R., et al. (2019). Exome capture for variant discovery and analysis in barley. *Barley: Methods Protoc.* 1900, 283–310. doi: 10.1007/978-1-4939-8944-7_18

Bharadwaj, C., Tripathi, S., Soren, K. R., Thudi, M., Singh, R. K., Sheoran, S., et al. (2021). Introgression of "QTL-hotspot" region enhances drought tolerance and grain yield in three elite chickpea cultivars. *Plant Genome* 14 (1), e20076. doi: 10.1002/tpg2.20076

Bhat, J. A., Yu, D., Bohra, A., Ganie, S. A., and Varshney, R. K. (2021). Features and applications of haplotypes in crop breeding. *Commun. Biol.* 4 (1), 1266. doi: 10.1038/s42003-021-02782-y

Bhattacharjee, S., Chakrabarty, A., Kora, D., and Roy, U. K. (2023). Hydrogen peroxide induced antioxidant-coupled redox regulation of germination in rice: redox metabolic, transcriptomic and proteomic evidences. *J. Plant Growth Regulation* 42 (2), 1084–1106. doi: 10.1007/s00344-022-10615-3

Borrell, A. K., Mullet, J. E., George-Jaeggli, B., van Oosterom, E. J., Hammer, G. L., Klein, P. E., et al. (2014). Drought adaptation of stay-green sorghum is associated with canopy development, leaf anatomy, root growth, and water uptake. *J. Exp. Botany* 65 (21), 6251–6263. doi: 10.1093/jxb/eru232

Cao, M., Zheng, L., Li, J., Mao, Y., Zhang, R., Niu, X., et al. (2022). Transcriptomic profiling suggests candidate molecular responses to waterlogging in cassava. *PloS One* 17 (1), e0261086. doi: 10.1371/journal.pone.0261086

Chen, K., Wang, Y., Zhang, R., Zhang, H., and Gao, C. (2019). CRISPR/Cas genome editing and precision plant breeding in agriculture. *Annu. Rev. Plant Biol.* 70, 667–697. doi: 10.1146/annurev-arplant-050718-100049

Cobb, J. N., Biswas, P. S., and Platten, J. D. (2019). Back to the future: revisiting MAS as a tool for modern plant breeding. *Theor. Appl. Genet.* 132, 647–667. doi: 10.1007/s00122-018-3266-4

Cobb, J. N., Chen, C., Shi, Y., Maron, L. G., Liu, D., Rutzke, M., et al. (2021). Genetic architecture of root and shoot ionomes in rice (Oryza sativa L.). *Theor. Appl. Genet.* 134 (8), 2613–2637. doi: 10.1007/s00122-021-03848-5

Coutinho, I. D., Henning, L. M. M., Döpp, S. A., Nepomuceno, A., Moraes, L. A. C., Marcolino-Gomes, J., et al. (2018). Flooded soybean metabolomic analysis reveals important primary and secondary metabolites involved in the hypoxia stress response and tolerance. *Environ. Exp. Botany* 153, 176–187. doi: 10.1016/j.envexpbot.2018.05.018

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Science* 22 (11), 961–975. doi: 10.1016/j.tplants.2017.08.011

Das, A., Eldakak, M., Paudel, B., Kim, D.-W., Hemmati, H., Basu, C., et al. (2016). Leaf proteome analysis reveals prospective drought and heat stress response mechanisms in soybean. *BioMed. Res. Int.* 2016, 1–23. doi: 10.1155/2016/6021047

Ding, H., Mo, S., Qian, Y., Yuan, G., Wu, X., and Ge, C. (2020). Integrated proteome and transcriptome analyses revealed key factors involved in tomato (Solanum lycopersicum) under high temperature stress. *Food Energy Security* 9 (4), e239. doi: 10.1002/fes3.239

Dong, Z., and Chen, Y. (2013). Transcriptomics: advances and approaches. *Sci. China Life Sci.* 56, 960–967. doi: 10.1007/s11427-013-4557-2

Eathington, S. R., Crosbie, T. M., Edwards, M. D., Reiter, R. S., and Bull, J. K. (2007). Molecular markers in a commercial breeding program. *Crop Science* 47, S–154-S-63. doi: 10.2135/cropsci2007.04.0015IPBS

Ehsan, A., Naqvi, R. Z., Azhar, M., Awan, M. J. A., Amin, I., Mansoor, S., et al. (2023). Genome-wide analysis of WRKY gene family and negative regulation of ghWRKY25 and ghWRKY33 reveal their role in whitefly and drought stress tolerance in cotton. *Genes* 14 (1), 171. doi: 10.3390/genes14010171

Ereful, N. C., L-y, L., Greenland, A., Powell, W., Mackay, I., and Leung, H. (2020). RNA-seq reveals differentially expressed genes between two indica inbred rice genotypes associated with drought-yield QTLs. *Agronomy* 10 (5), 621. doi: 10.3390/agronomy10050621

Farooq, M., Naqvi, R. Z., Amin, I., Rehman, A. U., Asif, M., and Mansoor, S. (2023). Transcriptome diversity assessment of Gossypium arboreum (FDH228) leaves under control, drought and whitefly infestation using PacBio long reads. *Gene* 852, 147065. doi: 10.1016/j.gene.2022.147065

Fatiukha, A., Klymiuk, V., Peleg, Z., Saranga, Y., Cakmak, I., Krugman, T., et al. (2020). Variation in phosphorus and sulfur content shapes the genetic architecture and phenotypic associations within the wheat grain ionome. *Plant J.* 101 (3), 555–572. doi: 10.1111/tpj.14554

Gajek, K., Janiak, A., Korotko, U., Chmielewska, B., Marzec, M., and Szarejko, I. (2021). Whole exome sequencing-based identification of a novel gene involved in root hair development in barley (Hordeum vulgare L.). *Int. J. Mol. Sci.* 22 (24), 13411. doi: 10.3390/ijms222413411

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51 (6), 1044–1051. doi: 10.1038/s41588-019-0410-2

Glushkevich, A., Spechenkova, N., Fesenko, I., Knyazev, A., Samarskaya, V., Kalinina, N. O., et al. (2022). Transcriptomic reprogramming, alternative splicing and RNA methylation in potato (Solanum tuberosum L.) plants in response to potato virus Y infection. *Plants* 11 (5), 635. doi: 10.3390/plants11050635

Gorthy, S., Narasu, L., Gaddameedi, A., Sharma, H. C., Kotla, A., Deshpande, S. P., et al. (2017). Introgression of shoot fly (Atherigona soccata L. Moench) resistance QTLs into elite post-rainy season sorghum varieties using marker assisted backcrossing (MABC). *Front. Plant science* 8, 1494. doi: 10.3389/fpls.2017.01494

Großkinsky, D. K., Syaifullah, S. J., and Roitsch, T. (2018). Integration of multi-omics techniques and physiological phenotyping within a holistic phenomics approach to study senescence in model and crop plants. *J. Exp. Botany* 69 (4), 825–844. doi: 10.1093/jxb/erx333

Guo, H., Li, S., Min, W., Ye, J., and Hou, Z. (2019). Ionomic and transcriptomic analyses of two cotton cultivars (Gossypium hirsutum L.) provide insights into the ion balance mechanism of cotton under salt stress. *PloS One* 14 (12), e0226776. doi: 10.1371/journal.pone.0226776

Guo, T., Yang, J., Li, D., Sun, K., Luo, L., Xiao, W., et al. (2019). Integrating GWAS, QTL, mapping and RNA-seq to identify candidate genes for seed vigor in rice (Oryza sativa L.). *Mol. Breed.* 39, 1–16. doi: 10.1007/s11032-019-0993-4

Guo, X., Xin, Z., Yang, T., Ma, X., Zhang, Y., Wang, Z., et al. (2020). Metabolomics response for drought stress tolerance in chinese wheat genotypes (Triticum aestivum). *Plants* 9 (4), 520. doi: 10.3390/plants9040520

Hasan, M. M., Rafii, M. Y., Ismail, M. R., Mahmood, M., Rahim, H. A., Alam, M. A., et al. (2015). Marker-assisted backcrossing: a useful method for rice improvement. *Biotechnol. Biotechnol. Equipment* 29 (2), 237–254. doi: 10.1080/13102818.2014.995920

Huang, T.-K., and Puchta, H. (2021). Novel CRISPR/Cas applications in plants: from prime editing to chromosome engineering. *Transgenic Res.* 30, 529–549. doi: 10.1007/s11248-021-00238-x

Huang, X.-Y., and Salt, D. E. (2016). Plant ionomics: from elemental profiling to environmental adaptation. *Mol. Plant* 9 (6), 787–797. doi: 10.1016/j.molp.2016.05.003

Illarionova, K., Grigoryev, S., Shelenga, T., and Rantakaulio, T. (2020). "Metabolomics approach in digital assessment of fatty acids profile of cottonseed for biological activity improvement of cotton oil," in *IOP Conference Series: Materials Science and Engineering* (Philadelphia, United States: IOP Publishing).

Jan, N., Rather, AM-U-D, John, R., Chaturvedi, P., Ghatak, A., Weckwerth, W., et al. (2023). Proteomics for abiotic stresses in legumes: present status and future directions. *Crit. Rev. Biotechnol.* 43 (2), 171–190. doi: 10.1080/07388551.2021.2025033

Jiang, H., Li, X., Ma, L., Ren, Y., Bi, Y., and Prusky, D. (2022). Transcriptome sequencing and differential expression analysis of natural and BTH-treated wound healing in potato tubers (Solanum tuberosum L.). *BMC Genomics* 23 (1), 1–20. doi: 10.1186/s12864-022-08480-1

Jiang, S.-Y., and Ramachandran, S. (2010). Assigning biological functions to rice genes by genome annotation, expression analysis and mutagenesis. *Biotechnol. letters* 32, 1753–1763. doi: 10.1007/s10529-010-0377-7

Kang, Y. J., Lee, T., Lee, J., Shim, S., Jeong, H., Satyawan, D., et al. (2016). Translational genomics for plant breeding with the genome sequence explosion. *Plant Biotechnol. J.* 14 (4), 1057–1069. doi: 10.1111/pbi.12449

Kasinathan, T., Singaraju, D., and Uyyala, S. R. (2021). Insect classification and detection in field crops using modern machine learning techniques. *Inf. Process. Agriculture* 8 (3), 446–457. doi: 10.1016/j.inpa.2020.09.006

Kaur, P., and Gaikwad, K. (2017). From genomes to GENE-omes: exome sequencing concept and applications in crop improvement. *Front. Plant science* 8, 2164. doi: 10.3389/fpls.2017.02164

Kaur, S., Shamshad, M., Jindal, S., Kaur, A., Singh, S., and Kaur, S. (2022). RNA-seq-based transcriptomics study to investigate the genes governing nitrogen use efficiency in Indian wheat cultivars. *Front. Genet.* 13, 461. doi: 10.3389/fgene.2022.853910

Kavil, S., Otti, G., Bouvaine, S., Armitage, A., and Maruthi, M. N. (2021). PAL1 gene of the phenylpropanoid pathway increases resistance to the Cassava brown streak virus in cassava. *Virol. J.* 18, 1–10. doi: 10.1186/s12985-021-01649-2

Khan, A., Sovero, V., and Gemenet, D. (2016). Genome-assisted breeding for drought resistance. *Curr. Genomics* 17 (4), 330–342. doi: 10.2174/1389202917999160211101417

Kim, M.-S., Yang, J.-Y., Yu, J.-K., Lee, Y., Park, Y.-J., Kang, K.-K., et al. (2021). Breeding of high cooking and eating quality in rice by marker-assisted backcrossing (MABc) using KASP markers. *Plants* 10 (4), 804. doi: 10.3390/plants10040804

Komatsu, S., Mock, H.-P., Yang, P., and Svensson, B. (2013). Application of proteomics for improving crop protection/artificial regulation. *Front. Media SA;* 4, 522. doi: 10.3389/fpls.2013.00522

Krishnan, A., Guiderdoni, E., An, G., Hsing, Y.-i., Han, C.-d., Lee, M. C., et al. (2009). Mutant resources in rice for functional genomics of the grasses. *Plant Physiol.* 149 (1), 165–170. doi: 10.1104/pp.108.128918

Kumar, A., Dixit, S., Ram, T., Yadaw, R., Mishra, K., and Mandal, N. (2014). Breeding high-yielding drought-tolerant rice: genetic variations and conventional and molecular approaches. *J. Exp. Botany* 65 (21), 6265–6278. doi: 10.1093/jxb/eru363

Kwak, G.-H., and Park, N.-W. (2019). Impact of texture information on crop classification with machine learning and UAV images. *Appl. Sci.* 9 (4), 643. doi: 10.3390/app9040643

Lacomme, C. (2015). Strategies for altering plant traits using virus-induced gene silencing technologies. *Plant Gene Silencing: Methods Protoc.* 1287, 25–41. doi: 10.1007/978-1-4939-2453-0_2

Lei, L., Zheng, H., Bi, Y., Yang, L., Liu, H., Wang, J., et al. (2020). Identification of a major QTL and candidate gene analysis of salt tolerance at the bud burst stage in rice (Oryza sativa L.) using QTL-Seq and RNA-Seq. *Rice* 13, 1–14. doi: 10.1186/s12284-020-00416-1

Lekota, M., Muzhinji, N., and van der Waals, J. E. (2019). Identification of differentially expressed genes in tolerant and susceptible potato cultivars in response to Spongospora subterranea f. sp. subterranea tuber infection. *Plant Pathology* 68 (6), 1196–1206. doi: 10.1111/ppa.13029

Li, Y., Ruperao, P., Batley, J., Edwards, D., Khan, T., Colmer, T. D., et al. (2018). Investigating drought tolerance in chickpea using genome-wide association mapping and genomic selection based on whole-genome resequencing data. *Front. Plant Science* 9, 190. doi: 10.3389/fpls.2018.00190

Li, Y., Xiong, H., Zhang, J., Guo, H., Zhou, C., Xie, Y., et al. (2022). Genome-wide and exome-capturing sequencing of a gamma-ray-induced mutant reveals biased variations in common wheat. *Front. Plant Science* 12, 793496. doi: 10.3389/fpls.2021.793496

Li, Y.-h., Zhou, G., Ma, J., Jiang, W., Jin, L.-g., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32 (10), 1045–1052. doi: 10.1038/nbt.2979

Li, Z., Zhu, A., Song, Q., Chen, H. Y., Harmon, F. G., and Chen, Z. J. (2020). Temporal regulation of the metabolome and proteome in photosynthetic and photorespiratory pathways contributes to maize heterosis. *Plant Cell.* 32 (12), 3706–3722. doi: 10.1105/tpc.20.00320

Liu, H., Liu, L., Liang, D., Zhang, M., Jia, C., Qi, M., et al. (2021). SlBES1 promotes tomato fruit softening through transcriptional inhibition of PMEU1. *Iscience* 24 (8), 102926. doi: 10.1016/j.isci.2021.102926

Mahmood, M. A., Naqvi, R. Z., and Mansoor, S. (2022). Engineering crop resistance by manipulating disease susceptibility genes. *Mol. Plant* 15 (10), 1511–1513. doi: 10.1016/j.molp.2022.09.010

Mahmood, M. A., Naqvi, R. Z., Rahman, S. U., Amin, I., and Mansoor, S. (2023). Plant virus-derived vectors for plant genome engineering. *Viruses* 15 (2), 531. doi: 10.3390/v15020531

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, Ji, Ebana, K., et al. (2015). Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J.* 81 (1), 13–23. doi: 10.1111/tpj.12681

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K. K., et al. (2017). The pangenome of hexaploid bread wheat. *Plant J.* 90 (5), 1007–1013. doi: 10.1111/tpj.13515

Muthamilarasan, M., Singh, N. K., and Prasad, M. (2019). Multi-omics approaches for strategic improvement of stress tolerance in underutilized crop species: a climate change perspective. *Adv. Genet.* 103, 1–38. doi: 10.1016/bs.adgen.2019.01.001

Muthuramalingam, P., Krishnan, S. R., Pandian, S., Mareeswaran, N., Aruni, W., Pandian, S. K., et al. (2018). Global analysis of threonine metabolism genes unravel key players in rice to improve the abiotic stress tolerance. *Sci. Rep.* 8 (1), 9270. doi: 10.1038/s41598-018-27703-8

Naqvi, R. Z., Farooq, M., Naqvi, S. A. A., Siddiqui, H. A., Amin, I., Asif, M., et al. (2020). "Big data analytics and advanced technologies for sustainable agriculture," in *Handbook of Smart Materials, Technologies, and Devices*, vol. 40. (Springer Cham Switzerland: Applications of Industry), 1–27. doi: 10.1038/s41598-017-15963-9

Naqvi, R. Z., Siddiqui, H. A., Mahmood, M. A., Najeebullah, S., Ehsan, A., Azhar, M., et al. (2022). Smart breeding approaches in post-genomics era for developing climate-resilient food crops. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.972164

Naqvi, R. Z., Zaidi, S.-e., Akhtar, K. P., Strickler, S., Woldemariam, M., Mishra, B., et al. (2017). Transcriptomics reveals multiple resistance mechanisms against cotton leaf curl disease in a naturally immune cotton species, Gossypium arboreum. *Sci. Rep.* 7 (1), 15880.

Naqvi, R. Z., Zaidi, S.-e., Mukhtar, M. S., Amin, I., Mishra, B., Strickler, S., et al. (2019). Transcriptomic analysis of cultivated cotton Gossypium hirsutum provides insights into host responses upon whitefly-mediated transmission of cotton leaf curl disease. *PloS One* 14 (2), e0210011. doi: 10.1371/journal.pone.0210011

Ohyanagi, H., Yano, K., Yamamoto, E., and Kitazumi, A. (2022). *Plant Omics: Advances in Big Data Biology* (Wallingford, UK: CABI).

Olsen, T. K., and Baryawno, N. (2018). Introduction to single-cell RNA sequencing. *Curr. Protoc. Mol. Biol.* 122 (1), e57. doi: 10.1002/cpmb.57

Pasion, E. A., Misra, G., Kohli, A., and Sreenivasulu, N. (2023). Unraveling the genetics underlying micronutrient signatures of diversity panel present in brown rice through genome–ionome linkages. *Plant J.* 113 (4), 749–771. doi: 10.1111/tpj.16080

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., et al. (2021). Machine learning for large-scale crop yield forecasting. *Agric. Systems* 187, 103016. doi: 10.1016/j.agsy.2020.103016

Peng, B., Li, H., and Peng, X.-X. (2015). Functional metabolomics: from biomarker discovery to metabolome reprogramming. *Protein Cell.* 6 (9), 628–637. doi: 10.1007/s13238-015-0185-x

Perez-Fons, L., Bohorquez-Chaux, A., Irigoyen, M. L., Garceau, D. C., Morreel, K., Boerjan, W., et al. (2019). A metabolomics characterisation of natural variation in the resistance of cassava to whitefly. *BMC Plant Biol.* 19 (1), 1–14. doi: 10.1186/s12870-019-2107-1

Pradhan, S. K., Pandit, E., Nayak, D. K., Behera, L., and Mohapatra, T. (2019). Genes, pathways and transcription factors involved in seedling stage chilling stress tolerance in indica rice through RNA-Seq analysis. *BMC Plant Biol.* 19 (1), 1–17. doi: 10.1186/s12870-019-1922-8

Pretorius, C. J., Tugizimana, F., Steenkamp, P. A., Piater, L. A., and Dubery, I. A. (2021). Metabolomics for biomarker discovery: Key signatory metabolic profiles for the identification and discrimination of oat cultivars. *Metabolites* 11 (3), 165. doi: 10.3390/metabo11030165

Rai, K. K. (2022). Integrating speed breeding with artificial intelligence for developing climate-smart crops. *Mol. Biol. Rep.* 49 (12), 11385–11402. doi: 10.1007/s11033-022-07769-4

Rama Reddy, N. R., Ragimasalawada, M., Sabbavarapu, M. M., Nadoor, S., and Patil, J. V. (2014). Detection and validation of stay-green QTL in post-rainy sorghum involving widely adapted cultivar, M35-1 and a popular stay-green genotype B35. *BMC Genomics* 15, 1–16. doi: 10.1186/1471-2164-15-909

Ramesh, S., and Vydeki, D. (2019). Application of machine learning in detection of blast disease in South Indian rice crops. *J. Phytol.* 11 (1), 31–37. doi: 10.25081/jp.2019.v11.5476

Raza, A. (2020). Metabolomics: a systems biology approach for enhancing heat stress tolerance in plants. *Plant Cell Rep.* 41, 741–763. doi: 10.1007/s00299-020-02635-8

Razzaq, A., Sadia, B., Raza, A., Khalid Hameed, M., and Saleem, F. (2019). Metabolomics: A way forward for crop improvement. *Metabolites* 9 (12), 303. doi: 10.3390/metabo9120303

Safavi-Rizi, V., Herde, M., and Stöhr, C. (2020). RNA-Seq reveals novel genes and pathways associated with hypoxia duration and tolerance in tomato root. *Sci. Rep.* 10 (1), 1–17. doi: 10.1038/s41598-020-57884-0

Sahoo, R. N., Viswanathan, C., Kumar, M., Bhugra, S., Karwa, S., Misra, T., et al. (2023). "High-Throughput Phenomics of Crops for Water and Nitrogen Stress," in *Translating Physiological Tools to Augment Crop Breeding* (Singapore: Springer), 291–310.

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., et al. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* 165 (3), 1120–1132. doi: 10.1104/pp.114.241521

Shami, A. A., Akhtar, M. T., Mumtaz, M. W., Mukhtar, H., Tahir, A., Shahzad-ul-Hussan, S., et al. (2023). NMR-based metabolomics: A new paradigm to unravel defense-related metabolites in insect-resistant cotton variety through different multivariate data analysis approaches. *Molecules* 28 (4), 1763. doi: 10.3390/molecules28041763

Sharma, S. K., McLean, K., Colgan, R. J., Rees, D., Young, S., Sønderkær, M., et al. (2021). Combining conventional QTL analysis and whole-exome capture-based bulk-segregant analysis provides new genetic insights into tuber sprout elongation and dormancy release in a diploid potato population. *Heredity* 127 (3), 253–265. doi: 10.1038/s41437-021-00459-0

Shokat, S., Großkinsky, D. K., Singh, S., and Liu, F. (2023). The role of genetic diversity and pre-breeding traits to improve drought and heat tolerance of bread wheat at the reproductive stage. *Food Energy Secur.* 12, e478. doi: 10.1002/fes3.478

Shokat, S., Sehgal, D., Vikram, P., Liu, F., and Singh, S. (2020). Molecular markers associated with agro-physiological traits under terminal drought conditions in bread wheat. *Int. J. Mol. Sci.* 21 (9), 3156. doi: 10.3390/ijms21093156

Singh, S., Jighly, A., Sehgal, D., Burgueño, J., Joukhadar, R., Singh, S., et al. (2021). Direct introgression of untapped diversity into elite wheat lines. *Nat. Food.* 2 (10), 819–827. doi: 10.1038/s43016-021-00380-z

Singh, S., Vikram, P., Sehgal, D., Burgueño, J., Sharma, A., Singh, S. K., et al. (2018). Harnessing genetic potential of wheat germplasm banks through impact-oriented-prebreeding for future food and nutritional security. *Sci. Rep.* 8 (1), 1–11. doi: 10.1038/s41598-018-30667-4

Siriwan, W., Vannatim, N., Chaowongdee, S., Roytrakul, S., Charoenlappanit, S., Pongpamorn, P., et al. (2023). Integrated proteomic and metabolomic analysis of cassava cv. Kasetsart 50 infected with Sri Lankan cassava mosaic virus. *Agronomy* 13 (3), 945. doi: 10.3390/agronomy13030945

Song, J., Pei, W., Wang, N., Ma, J., Xin, Y., Yang, S., et al. (2022). Transcriptome analysis and identification of genes associated with oil accumulation in upland cotton. *Physiologia Plantarum* 174 (3), e13701. doi: 10.1111/ppl.13701

Su, W.-H. (2020). Advanced machine learning in point spectroscopy, RGB-and hyperspectral-imaging for automatic discriminations of crops and weeds: A review. *Smart Cities* 3 (3), 767–792. doi: 10.3390/smartcities3030039

Swamy, B. P. M., Ahmed, H. U., Henry, A., Mauleon, R., Dixit, S., Vikram, P., et al. (2013). Genetic, physiological, and gene expression analyses reveal that multiple QTL enhance yield of rice mega-variety IR64 under drought. *PloS One* 8 (5), e62795. doi: 10.1371/journal.pone.0062795

Templer, S. E., Ammon, A., Pscheidt, D., Ciobotea, O., Schuy, C., McCollum, C., et al. (2017). Metabolite profiling of barley flag leaves under drought and combined heat and drought stress reveals metabolic QTLs for metabolites associated with antioxidant defense. *J. Exp. botany* 68 (7), 1697–1713. doi: 10.1093/jxb/erx038

Tiwari, J. K., Buckseth, T., Devi, S., Varshney, S., Sahu, S., Patil, V. U., et al. (2020). Physiological and genome-wide RNA-sequencing analyses identify candidate genes in a nitrogen-use efficient potato cv. Kufri Gaurav. *Plant Physiol. Biochem.* 154, 171–183. doi: 10.1016/j.plaphy.2020.05.041

Varshney, R. K., Bohra, A., Yu, J., Graner, A., Zhang, Q., and Sorrells, M. E. (2021). Designing future crops: genomics-assisted breeding comes of age. *Trends Plant Science* 26 (6), 631–649. doi: 10.1016/j.tplants.2021.03.010

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends Plant Science* 10 (12), 621–630. doi: 10.1016/j.tplants.2005.10.004

Varshney, R. K., Saxena, R. K., Upadhyaya, H. D., Khan, A. W., Yu, Y., Kim, C., et al. (2017). Whole-genome resequencing of 292 pigeonpea accessions identifies genomic regions associated with domestication and agronomic traits. *Nat. Genet.* 49 (7), 1082–1088. doi: 10.1038/ng.3872

Viana, V. E., Pegoraro, C., Busanello, C., and Costa de Oliveira, A. (2019). Mutagenesis in rice: the basis for breeding a new super plant. *Front. Plant science* 10, 1326. doi: 10.3389/fpls.2019.01326

Wang, N., Long, T., Yao, W., Xiong, L., Zhang, Q., and Wu, C. (2013). Mutant resources for the functional analysis of the rice genome. *Mol. Plant* 6 (3), 596–604. doi: 10.1093/mp/sss142

Wen, J., Jiang, F., Weng, Y., Sun, M., Shi, X., Zhou, Y., et al. (2019). Identification of heat-tolerance QTLs and high-temperature stress-responsive genes through conventional QTL mapping, QTL-seq and RNA-seq in tomato. *BMC Plant Biol.* 19, 1–17. doi: 10.1186/s12870-019-2008-3

Wu, C., Ding, X., Ding, Z., Tie, W., Yan, Y., Wang, Y., et al. (2019). The class III peroxidase (POD) gene family in cassava: identification, phylogeny, duplication, and expression. *Int. J. Mol. Sci.* 20 (11), 2730. doi: 10.3390/ijms20112730

Xi, W., Hao, C., Li, T., Wang, H., and Zhang, X. (2023). Transcriptome analysis of roots from wheat (Triticum aestivum L.) varieties in response to drought stress. *Int. J. Mol. Sci.* 24 (8), 7245. doi: 10.3390/ijms24087245

Xie, J., Guo, G., Wang, Y., Hu, T., Wang, L., Li, J., et al. (2020). A rare single nucleotide variant in Pm5e confers powdery mildew resistance in common wheat. *New Phytologist* 228 (3), 1011–1026. doi: 10.1111/nph.16762

Xiong, H., Guo, H., Fu, M., Xie, Y., Zhao, L., Gu, J., et al. (2023). A large-scale whole-exome sequencing mutant resource for functional genomics in wheat. *Plant Biotechnol. J.* 21 (10), 2047–2056. doi: 10.1111/pbi.14111

Xu, Y., Zhang, X.-Q., Harasymow, S., Westcott, S., Zhang, W., and Li, C. (2018). Molecular marker-assisted backcrossing breeding: an example to transfer a thermostable β-amylase gene from wild barley. *Mol. Breeding* 38, 1–9. doi: 10.1007/s11032-018-0828-8

Yang, Z., Wang, Z., Yang, C., Yang, Z., Li, H., and Wu, Y. (2019). Physiological responses and small RNAs changes in maize under nitrogen deficiency and resupply. *Genes Genomics* 41, 1183–1194. doi: 10.1007/s13258-019-00848-0

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., et al. (2019). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* 17 (5), 881–892. doi: 10.1111/pbi.13022

Yu, X., Wang, T., Zhu, M., Zhang, L., Zhang, F., Jing, E., et al. (2019). Transcriptome and physiological analyses for revealing genes involved in wheat response to endoplasmic reticulum stress. *BMC Plant Biol.* 19 (1), 1–22. doi: 10.1186/s12870-019-1798-7

Zaidi, S., Naqvi, R. Z., Asif, M., Strickler, S., Shakir, S., Shafiq, M., et al. (2020). Molecular insight into cotton leaf curl geminivirus disease resistance in cultivated cotton (Gossypium hirsutum). *Plant Biotechnol. J.* (Cham Switzerland: Springer) 18 (3), 691–706. doi: 10.1111/pbi.13236

Zhang, C., Liu, S., Li, X., Zhang, R., and Li, J. (2022). Virus-induced gene editing and its applications in plants. *Int. J. Mol. Sci.* 23 (18), 10202. doi: 10.3390/ijms231810202

Zhang, J., Mei, H., Lu, H., Chen, R., Hu, Y., and Zhang, T. (2022). Transcriptome time-course analysis in the whole period of cotton fiber development. *Front. Plant Science* 13, 804. doi: 10.3389/fpls.2022.864529

Zhang, H., Zhang, J., Xu, Q., Wang, D., Di, H., Huang, J., et al. (2020). Identification of candidate tolerance genes to low-temperature during maize germination by GWAS and RNA-seq approaches. *BMC Plant Biol.* 20, 1–17. doi: 10.1186/s12870-020-02543-9

Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50 (2), 278–284. doi: 10.1038/s41588-018-0041-z

Zhao, X., and Niu, Y. (2022). The combination of conventional QTL analysis, bulked-segregant analysis, and RNA-sequencing provide new genetic insights into maize mesocotyl elongation under multiple deep-seeding environments. *Int. J. Mol. Sci.* 23 (8), 4223. doi: 10.3390/ijms23084223

Zhao, J., Sauvage, C., Bitton, F., and Causse, M. (2022). Multiple haplotype-based analyses provide genetic and evolutionary insights into tomato fruit weight and composition. *Horticulture Res.* 9, uhab009. doi: 10.1093/hr/uhab009

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33 (4), 408–414. doi: 10.1038/nbt.3096

Zhu, H., Li, C., and Gao, C. (2020). Applications of CRISPR–Cas in agriculture and plant biotechnology. *Nat. Rev. Mol. Cell Biol.* 21 (11), 661–677. doi: 10.1038/s41580-020-00288-9

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fpls.2023.1305547

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Jindong Liu,
Chinese Academy of Agricultural Sciences, China
Jian Ma,
Sichuan Agricultural University, China

*CORRESPONDENCE
Bo Feng
fengbo@cib.ac.cn

RECEIVED 02 October 2023
ACCEPTED 08 December 2023
PUBLISHED 08 January 2024

CITATION
Jiang C, Xu Z, Fan X, Zhou Q, Ji G, Liao S, Wang Y, Ma F, Zhao Y, Wang T and Feng B (2024) Genetic dissection of major QTL for grain number per spike on chromosomes 5A and 6A in bread wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 14:1305547. doi: 10.3389/fpls.2023.1305547

Genetic dissection of major QTL for grain number per spike on chromosomes 5A and 6A in bread wheat (*Triticum aestivum* L.)

Cheng Jiang[1,2,3], Zhibin Xu[1], Xiaoli Fan[1], Qiang Zhou[1], Guangsi Ji[1,3], Simin Liao[1,3], Yanlin Wang[1,3], Fang Ma[1,3], Yun Zhao[2], Tao Wang[1,4] and Bo Feng[1]*

[1]Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu, China, [2]College of Life Sciences, Sichuan University, Chengdu, China, [3]University of Chinese Academy of Sciences, Beijing, China, [4]The Innovative of Seed Design, Chinese Academy of Sciences, Beijing, China

Grain number per spike (GNS) is a crucial component of grain yield and plays a significant role in improving wheat yield. To identify quantitative trait loci (QTL) associated with GNS, a recombinant inbred line (RIL) population derived from the cross of Zhongkemai 13F10 and Chuanmai 42 was employed to conduct QTL mapping across eight environments. Based on the bulked segregant exome sequencing (BSE-Seq), genomic regions associated with GNS were detected on chromosomes 5A and 6A. According to the constructed genetic maps, two major QTL *QGns.cib-5A* (LOD = 4.35–8.16, PVE = 8.46–14.43%) and *QGns.cib-6A* (LOD = 3.82–30.80, PVE = 5.44–12.38%) were detected in five and four environments, respectively. *QGns.cib-6A* is a QTL cluster for other seven yield-related traits. *QGns.cib-5A* and *QGns.cib-6A* were further validated using linked Kompetitive Allele Specific PCR (KASP) markers in different genetic backgrounds. *QGns.cib-5A* exhibited pleiotropic effects on productive tiller number (PTN), spike length (SL), fertile spikelet number per spike (FSN), and ratio of grain length to grain width (GL/GW) but did not significantly affect thousand grain weight (TGW). Haplotype analysis revealed that *QGns.cib-5A* and *QGns.cib-6A* were the targets of artificial selection during wheat improvement. Candidate genes for *QGns.cib-5A* and *QGns.cib-6A* were predicted by analyzing gene annotation, spatiotemporal expression patterns, and orthologous and sequence differences. These findings will be valuable for fine mapping and map-based cloning of genes underlying *QGns.cib-5A* and *QGns.cib-6A*.

KEYWORDS

QTL mapping, BSE-Seq, grain number per spike, haplotype analysis, wheat

Introduction

Wheat (*Triticum aestivum* L.) is a vital crop that provides a substantial portion of the world's food. However, as the world population continues to grow, the demand for food is increasing. Despite the current annual growth rate of wheat production reaching 0.9%, it falls short of the required annual growth rate of approximately 2.4% needed to sustain the world population by 2050 (Ray et al., 2013; Gao, 2021). As a result, enhancing the yield potential has become a fundamental objective in wheat breeding. Grain yield is a complex quantitative trait mainly determined by three factors: spike number per unit area, thousand grain weight (TGW), and grain number per spike (GNS). Therefore, revealing the genetic factors underlying GNS is essential to improve yield potential.

The genetic regulatory pathways governing architecture of the inflorescence play a crucial role in determining GNS in wheat. Generally, spike development can be divided into three main phases: the duration of the flowering transition; initiation, distribution, and termination of spikelet meristem (SM); formation and generation of floret meristem (FM) (Luo et al., 2023). During the flowering transition period, several widely recognized genes involved in flowering time participate in regulating the timing of inflorescence meristem (IM) differentiation and the initiation of spikelet and floret development. During the vernalization-induced flowering process, *VERNALIZATION 1* (*VRN1*) serves as a central regulatory gene, playing a crucial role in maintaining IM activity and controlling SM characteristics. Similar to VRN1, FRUITFULL2 (FUL2) and FRUITFULL3 (FUL3) redundantly facilitate the transition from SAM to IM (Li et al., 2019). The photoperiod gene *Photoperiod-1* (*Ppd-1*) in wheat influences inflorescence structure. Insufficient Ppd-D1 leads to the formation of paired spikelet and an increase in grain count under short sunlight conditions (Boden et al., 2015). VRN3/TaFT1, a homolog of Flowering Locus T (FT) in *Arabidopsis* and the Heading date 3a (Hd3a) in rice, interacts with the transcription factor TaFDL2 to activate *VRN1* (Yan et al., 2006; Li and Dubcovsky, 2008). Additionally, VRN1 and Ppd-1 positively regulate *VRN3*/*TaFT1* and *TaFT2*, whereas VRN2 is known to act as a transcriptional inhibitor of these genes (Yan et al., 2004; Chen and Dubcovsky, 2012; Shaw et al., 2019). TaFT2 controls the initiation and quantification of spikelets. *PHYTOCHROME C* (*PHYC*) acts as an upstream regulatory factor, activating *Ppd-1* and *VRN3*/*TaFT1*. The TCP transcription factor TEOSINTE BRANCHED 1 (TaTB1) inhibits spike formation (Dixon et al., 2018).

The transition from IM to SM is crucial for establishing the inflorescence structure in Gramineae plants during spikelet initiation, distribution, and termination. Overexpression of wheat *AGAMOUS-LIKE6* (*TaAGL6*) affects the expression of meristem-active genes like *FUL2* and *TaMADS55*, resulting in a significant increase in the number of spikelets and grains per spike (Kong et al., 2021). Wheat FRIZY PANICLE (WFZP), a member of the class II AP2/ERF transcription factor (TF), directly activates *VRN1-A* and *HOMEOBOX4* (*TaHOX4*)*-A*. In addition, *WFZP* also acts as an inhibitor of the spikelet formation gene *BARREN STALK1* (*TaBA1*), exerting a dual effect (Poursarebani et al., 2015; Du et al., 2021; Li et al., 2021). The microRNA156 (miR156)-SPL module is crucial in initiating SM development during wheat spike development. MiR156 regulates SPL family genes, including *TaSPL3*/*17* in wheat. TaSPL3/17 interacts with DWARF53 (TaD53) to regulate the expression of genes *TaTB1* and *TaBA1*, which are involved in the differentiation of spikelet meristem and floral meristem. This interaction ultimately affects wheat spikelet development (Liu et al., 2017). The miR172-AP2 module also plays a critical role in the proper development of spikelets (Debernardi et al., 2017; Zhong et al., 2021). The *q*/*ap2l5* mutant exhibits a significant reduction in spikelet number, which can be attributed to the premature transformation of spikelet meristem into terminal spikelet (Debernardi et al., 2020).

The interactions among MADS, SPL, TCP, and AP2 TFs play an essential role in promoting or maintaining the characteristics of SM and FM, which significantly influence the development of wheat floret. The E-class *SEP* genes are primarily responsible for regulating the floral organs' development (Pelaz et al., 2000; Ditta et al., 2004). Upregulation *TaVRT2*, a MADS-box gene belonging to the *SHORT VEGETATIVE PHASE* (*SVP*) branch, causes the downregulation of *TaSEP1*. As a result, the transformation of spikelet into floret is delayed, resulting in an increased number of basal spikelets (Backhaus et al., 2022). The SQUAMOSA proteins VRN1 and FUL2 function as repressors of the *SVP* branch MADS box genes, such as *TaVRT2*, *TaSVP1*, and *TaSVP3*. These proteins stimulate the formation of small flowers following the transition to flowering (Li et al., 2019; Li et al., 2021; Liu et al., 2021). Therefore, the downregulation of SQUAMOSA protein for *SVP* gene expression is essential to promoting *SEP* gene expression and ensuring normal flower development (Li et al., 2021; Backhaus et al., 2022). Upregulation of the miR156 target gene *TaSPL13* leads to an increased production of small flowers and grains per spike in wheat (Li et al., 2020). *Q*/*AP2L5* and *AP2L2* redundantly recognize and prevent small flowers from degenerating into glumes through miR172-guided mechanisms (Debernardi et al., 2017; Debernardi et al., 2020).

Like other traits related to yield, GNS is a quantitative trait influenced by both genetic and environmental factors. As a result, researchers have preliminarily focused on mapping quantitative trait loci (QTL) in various genetic or natural populations of wheat. Up to now, numerous QTL associated with GNS have been identified across 21 chromosomes in previous studies (Börner et al., 2002; Peng et al., 2003; Huang et al., 2004; Quarrie et al., 2005; Liu et al., 2006; Kumar et al., 2007; Cuthbert et al., 2008; Wang et al., 2009; McIntyre et al., 2010; Wang et al., 2011; Blanco et al., 2012; Rustgi et al., 2013; Azadi et al., 2014; Gao et al., 2015; Zhang et al., 2016; Roncallo et al., 2017; Guan et al., 2018; Liu et al., 2019; Hu et al., 2020; Mizuno et al., 2021; Qiao et al., 2022; Hu et al., 2023). However, few major QTL have been found that can be detected in multiple environments and validated in different genetic backgrounds, hindering their utilization in breeding programs. Therefore, it is essential to identify and validate the novel QTL/genes associated with GNS.

In the present study, we utilized bulked segregant exome sequencing (BSE-Seq) and linkage analysis to identify QTL that

control GNS. The major QTL were subsequently validated in different genetic backgrounds, and potential candidate genes were predicted. Additionally, an analysis of the haplotypes of the major QTL was conducted.

Materials and methods

Plant materials and field trials

Three genetic populations obtained through the single-seed descent method as well as a natural population were employed in this study. They were (1) a recombinant inbred line (RIL) population (13CM, 316 F_7 lines) derived from the cross of Zhongkemai 13F10 (ZKM13F10) and Chuanmai 42 (CM42); (2) an F_2 population (CZ5782, 184 individuals) derived from the cross of Chuanmai 104 (CM104) and ZM5782; (3) an F_2 population (CS352, 126 individuals) derived from the cross of CM104 and SH352; and (4) a natural population containing 321 wheat accessions, including 59 widely grown cultivars during the last two decades and 262 accessions of Chinese wheat mini-core collection (88 modern cultivars, 17 introduced cultivars, and 157 landraces) (Li et al., 2022). The 13CM population was used to construct genetic map and detect QTL; CZ5782 and CS352 were used to validate the target QTL in different genetic backgrounds, and the natural population was used for haplotype analysis.

ZKM13F10 (ZKM138/PW18) is a stable breeding line selected by our lab characterized by high GNS. CM42 (Syncd768/SW3243//Chuan6415) is a core cultivar that has been used as one of the parents to develop more than 50 new cultivars in China. It possesses desirable yield-related traits including high grain weight and wide adaptability. CM104 is a cultivar derived from CM42 and inherits its major elite traits (including high grain weight and long spike). SH352 and ZM5782 are wheat lines to construct populations used for validating the major QTL.

The 13CM population and its parents were cultivated in eight different environments: Shuangliu (103°52′E, 30°34′N) during the 2017–2018, 2018–2019, 2019–2020, and 2020–2021 growing seasons (referred to as E1, E3, E5, and E7, respectively), and Shifang (104°11′E, 31°6′N) during the 2017–2018, 2018–2019, 2019–2020, and 2020–2021 seasons (referred to as E2, E4, E6, and E8, respectively). CZ5782 and CS352 individuals were cultivated in Shifang during the 2021–2022 growing season. Each plot had two rows. The row length and row spacing were 1.2 m and 0.2 m, respectively, and each row sowed 12 seeds. At sowing time, the fertilizer (N: 25%, P_2O_5: 10%, K_2O: 10%) was applied with 450 kg/ha. The local standard practices were applied in field management.

Phenotypic evaluation and statistical analysis

At maturity, eight plants from each line of 13CM, as well as the parents, were randomly selected for phenotypic evaluation. Traits including plant height (PHT), productive tiller number (PTN), spike length (SL), fertile spikelet number (FSN), and GNS were measured manually. The average values of these traits from the eight selected plants in each line were utilized for statistical analysis. Additionally, after air-drying, grain length (GL), grain width (GW), the ratio of GL to GW (GL/GW), and thousand grain weight (TGW) were measured. The spike compactness (SC) was calculated by dividing FSN by SL. GNFS was calculated by dividing GNS by FSN. The detailed method was conducted as described previously (Ji et al., 2021).

Descriptive statistics, Pearson's correlation analysis, normal distribution, and Student's *t*-test were carried out using SPSS v24.0 (SPSS, Chicago, USA). The QTL IciMapping v4.2 software (https://isbreeding.caas.cn/rj/qtllcmapping/) was used to calculate the broad-sense heritability (H^2) and the best linear unbiased estimation (BLUE) dataset for each trait. OriginPro v2019 (https://www.originlab.com/) was employed to create the histogram distribution, scatter plot, and box plot. The Pearson's correlation coefficients were utilized to examine the correlations between GNS and the other traits. Furthermore, by considering the genotypes of the flanking markers, lines harboring different alleles were compared using Student's *t*-test, with a significance level set at $P < 0.05$.

BSE-Seq analysis

The high-quality genomic DNA from 13CM lines and the parents was extracted by a modified hexadecyltrimethylammonium bromide (CTAB) method. Based on the phenotypic data obtained in E1–E6, lines in each environment were rearranged from low to high. To construct extreme mixing pools, 30 lines within each of two tails with stable phenotype in at least four of the six environments were selected. Two pools (GNS-H and GNS-L) were bulked using an equal amount (1 μg) of DNA from the selected 30 individuals. The two pools and the parents were utilized for BSE-Seq analysis performed by Bioacme Biotechnology Co., Ltd. (Wuhan, China).

The raw data from this study have been deposited in the Genome Sequence Archive (Chen et al., 2021) at the National Genomics Data Center (CNCB-NGDC Members and Partners, 2022), which is a part of the China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA008821 for ZKM13F10 and CM42, CAR009113 for GNS-H and GNS-L). These datasets are publicly accessible and can be found at https://ngdc.cncb.ac.cn/gsa. The processing of raw data was performed according to the previously method (Ji et al., 2023). In this study, two methods Euclidean distance (ED) and Δ(SNP-index) were employed to identify SNP and InDel sites between the paired pools. The detailed analytical method was described previously (Yu et al., 2022).

Development of molecular markers

Based on the BSE-Seq data, SNP/InDel in the associated genomic regions between the parents and extreme pools were converted to

develop Kompetitive Allele-specific PCR (KASP) markers for genetic map construction. Common primers were designed from Triticae Multi-omics Center (http://202.194.139.32/). FAM and HEX probe sequences were added to the 5' end of primers. The KASP genotype identification was performed in the QuantStudio™ 3 Real-Time PCR system designed by Thermo Fisher Scientific, with a reaction mixture containing 5 μL 2× main mixture, 0.8 μL primer mixture, 3 μL ddH_2O, and 2 μL DNA template (50 ng/mL–100 ng/mL). The conditions and procedures for touchdown PCR was referred to Yu et al. (2022).

Genetic map construction and QTL detection

The genetic map was constructed by JoinMap v4.1, and the QTL was detected by QTL IciMapping v4.2 (Meng et al., 2015). Markers that co-localized with others and had a missing rate more than 20% were discarded. The maximum likelihood mapping algorithm and Kosambi's mapping function were utilized to establish marker order and calculate marker distance. QTL detection in each environment was conducted using the QTL IciMapping v4.2 software based on the Inclusive Composite Interval Mapping (ICIM) method in the Biparental Population (BIP) module, and the LOD threshold was set as 2.5. The interaction of QTL×environment (QE) was performed using QTL IciMapping v4.2 according to Multi-Environment Trails module (LOD = 2.5, PIN = 0.001, and step = 4 cM). QTL repeatedly identified in at least three environments were treated as stable. Moreover, QTL explaining more than 10% of the phenotypic variation was considered as major. Confidence intervals were estimated by the position ± 1 LOD. QTL with overlapping confidence intervals were considered equivalent and named according to the international genetic naming rules (http://wheat.pw.usda.gov/ggpages/wgc/98/Intro.htm), where 'cib' represents 'Chengdu Institute of Biology'.

Haplotype analysis

Haplotypes at the crucial regions of the major QTL were analyzed based on the resequencing data of 145 landmark cultivars in China (http://wheat.cau.edu.cn/Wheat_SnpHub_Portal/). Subsequently, a natural population comprising 321 wheat accessions was used to conduct haplotype analysis. These accessions were planted in Shifang during the 2022–2023 growing season. The planting and phenotypic evaluation were conducted following the same protocols as the 13CM lines.

Prediction of candidate genes

Based on the mapping results, the physical positions of the flanking markers were converted from IWGSC RefSeq v1.0 to v2.1 using the Triticae Multi-omics Center (Zhu et al., 2021). The annotation and function of the genes located between the flanking markers were analyzed using Uniport (https://www.uniprot.org/). The expression pattern of the candidate genes was obtained from expVIP (http://www.wheat-expression.com/). These expression data were normalized using the ZeroToOne method and further presented in the HeatMap drawn by TBtools (Chen et al., 2020). The orthologues from rice (*Oryza sativa* L. *Japonica* group) and *Arabidopsis thaliana* were identified using Ensembl Plants (https://plants.ensembl.org/Triticum_aestivum/Info/Index). The functional information of these orthologues was obtained from China Rice Data Center (https://www.ricedata.cn/) for rice orthologues and *tair* (https://www.arabidopsis.org/) for *Arabidopsis* orthologues. In addition, based on the BSE-Seq data, nonsynonymous SNPs present in the exon regions of genes within the target regions were collected.

Results

Phenotypic performance

The GNS of ZKM13F10 was higher than that of CM42 in most environments. Significant differences in GNS between ZKM13F10 and CM42 were observed in E1, E4, E7 and the BLUE dataset ($P < 0.05$ or $P < 0.01$) (Table 1). In the 13CM population, GNS showed extensive and significant variation. Based on the BLUE dataset, the range of GNS variation was 42.86–72.47. The estimated value of H^2 of GNS was 0.83, indicating that GNS was mainly controlled by genetic factors. The continuous distribution of GNS across eight environments and the BLUE dataset showed that it is a typical quantitative trait controlled by multiple genes (Supplementary Figure 1 and Table 1). In multiple environments, the significant Pearson correlation of GNS ranged from 0.27 to 0.99 ($P < 0.001$) (Supplementary Figure 1).

Correlation analysis between GNS and other yield-related traits

The Pearson's correlation between GNS and other yield-related traits was evaluated using the BLUE dataset (Figure 1). Significant and negative correlations were detected between GNS and GL, GW, GL/GW, TGW, and PHT ($P < 0.01$ or $P < 0.001$). Moreover, GNS was significantly ($P < 0.001$) and positively correlated with FSN, GNFS, and SC. No significant correlation was observed between GNS and PTN ($r = -0.069$) or SL ($r = 0.022$). The correlation coefficient between GNS and GNFS was highest ($r = 0.83$).

BSE-Seq analyses

Based on the BSE-Seq data from four libraries, genomic regions associated with GNS were detected (Supplementary Table 1). After filtering, the numbers of clean reads in the four libraries were 77,283,512 (ZKM13F10), 91,988,190 (CM42), 119,966,894 (GNS-H), and 171,166,254 (GNS-L), respectively. This result indicates

TABLE 1 Phenotypic variation and heritability (H^2) of grain number per spike (GNS) of the parents and 13CM lines in eight environments and the BLUE dataset.

Env.	Parents		13CM lines					H^2
	ZKM13F10	CM42	Range	Mean ± SD	SK.	Ku.	CV (%)	
E1	77.67	49.67**	43.17–86.50	60.75 ± 6.21	0.40	0.92	10.2	0.83
E2	N	N	42.60–79.25	59.56 ± 5.88	0.15	0.14	9.6	
E3	68.00	65.33	42.30–81.09	60.85 ± 0.38	0.20	0.30	9.8	
E4	64.50	50.00*	43.45–80.84	59.46 ± 0.40	0.30	0.30	10.5	
E5	60.30	63.71	38.67–71.17	56.15 ± 0.37	−0.14	0.22	10.1	
E6	N	N	41.30–75.82	57.30 ± 0.41	−0.06	0.22	10.9	
E7	63.83	48.88**	39.32–76.16	56.84 ± 0.41	0.17	0.77	11.1	
E8	63.17	53.00	29.00–80.88	60.83 ± 0.47	−0.37	0.86	11.9	
BLUE	64.24	50.19*	42.86–72.47	58.52 ± 0.30	−0.03	0.27	7.9	

Env., environment; BLUE, best linear unbiased estimation; SD, standard deviation; SK., skewness; Ku., kurtosis; CV, coefficient of variation; N, the data were missed; H^2, broad-sense heritability; * and ** represent significant at $P < 0.05$ and $P < 0.01$.

that the data volume is sufficient for the subsequent analysis (Supplementary Table 2). Approximately 99.44% or higher of the captured reads were successfully mapped to the reference genome. The average sequencing depths ranged from 20.03× to 64.96×. Moreover, the coverage ≥5× varied from 60.38% to 78.15% in the four libraries, demonstrating high quality and adequate sequencing depth for BSE-Seq analysis. A total of 5,969,324 SNPs were identified in the dataset, and the number of SNPs per chromosome ranged from 58,929 to 630,245.

The ED and Δ(SNP-index) methods were used to detect genomic regions associated with GNS. Based on these results, genomic regions associated with GNS were detected on chromosomes 5A and 6A by ED and on chromosomes 4B, 5A, and 6A by Δ(SNP-index), respectively (Figure 2 and Supplementary Table 3). Specifically, the overlapping physical intervals detected by both methods were found in the range of 404.14 Mb–440.88 Mb on chromosome 5A and in 265.97 Mb–320.49 Mb on chromosome 6A.

Genetic map construction and QTL analysis

To confirm the preliminarily detected genomic regions associated with GNS, the polymorphic SNP sites within expanded regions (chr5A: 332.84 Mb–532.48 Mb; chr6A: 80.04 Mb–486.84 Mb) were converted into KASP markers (Supplementary Tables 4, 12). The phenotypic data evaluated in eight environments and the combined analysis (the BLUE dataset was set as an additional environment) were used for QTL mapping.

For chromosome 5A, 19 KASP markers were successfully developed to construct a genetic map with a length of 41.1 cM. According to this map, a major and stable additive QTL *QGns.cib-5A* was detected in five environments including the BLUE dataset (Figures 3A and 4). It explained 8.46%–14.43% of phenotypic variance, and the LOD values varied from 4.35 to 8.16. The favorable allele of *QGns.cib-5A* was contributed by ZKM13F10,

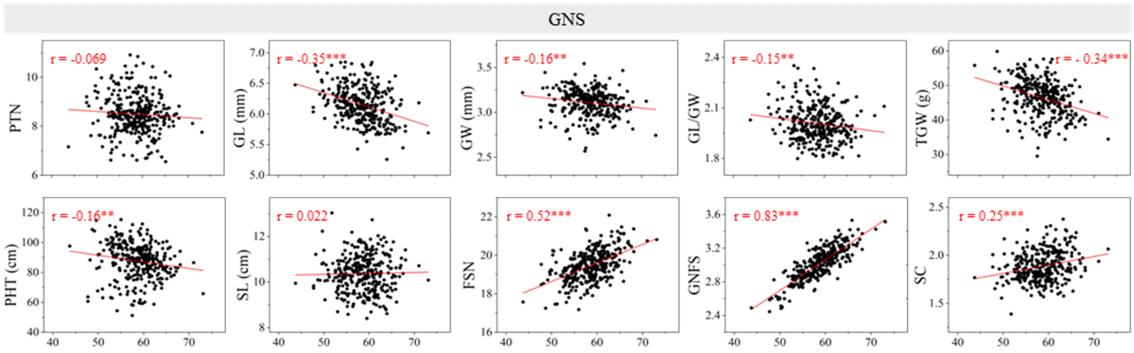


FIGURE 1
Coefficients of the pairwise Pearson's correlations between grain number per spike (GNS) and other yield-related traits in the 13CM population. The traits include productive tiller number (PTN), grain length (GL), grain width (GW), GL/GW, thousand grain weight (TGW), plant height (PHT), spike length (SL), fertile spikelet number per spike (FSN), grain number per fertile spikelet (GNFS), and spike compactness (SC) (the coefficient of the pairwise Pearson's correlations between GNS and FSN, GNFS have been published in Jiang et al., 2023). ** and *** represent significance at $P < 0.01$ and $P < 0.001$.

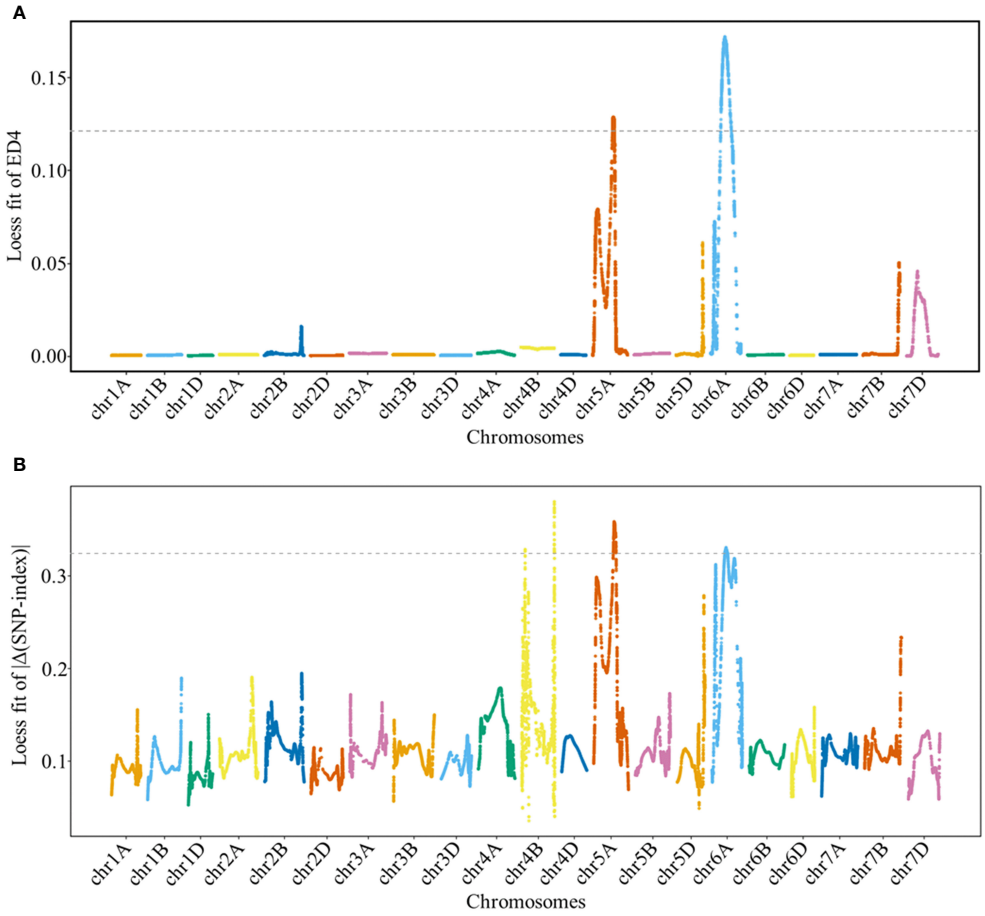


FIGURE 2
Locally weighted scatterplot smoothing (LOESS) fitting Manhattan plot for grain number per spike (GNS). Panels **(A, B)** show the LOESS fits of ED^4 and |Δ(SNP-index)| for GNS, respectively. The cutoff values for the two methods are indicated by the dotted lines, with threshold values of 0.1214 and 0.3243 for the LOESS fits of ED^4 and |Δ(SNP-index)|, respectively.

and this QTL was located in a 2.8-cM genetic interval between the markers *KASP12* and *KASP13* (Table 2).

For chromosome 6A, 19 KASP markers were developed and the genetic map spanned 20.4 cM in length. *QGns.cib-6A*, a major and stable additive QTL, was identified in E5, E7, E8, and the BLUE dataset (Figures 3B and 5A). It explained 5.44%–12.38% of phenotypic variance with the LOD values ranging from 3.82 to 30.80. The favorable allele of *QGns.cib-6A* was contributed by ZKM13F10. The QTL was located in a 0.2-cM genetic interval between the markers *KASP26* and *KASP27* (Table 2).

In addition, seven QTL were identified between the markers *KASP26* and *KASP27* on chromosome 6A (Supplementary Table 5). Three major and stable QTL (*QTgw.cib-6A*, *QGl.cib-6A*, and *QGw.cib-6A*) related to grain size and weight were detected (Figures 5F–H). *QTgw.cib-6A* was detected in five environments and the BLUE dataset and explained 10.26%–19.94% of phenotypic variance, with the LOD values ranging from 7.20 to 15.02. The *QGl.cib-6A* (LOD = 6.01–48.38; PVE = 8.68%–14.77%) was detected in four environments and the BLUE dataset. *QGw.cib-6A* (LOD = 5.70–12.48; PVE = 8.12%–16.74%) was detected in four environments and the BLUE dataset. The favorable alleles of *QTgw.cib-6A*, *QGl.cib-6A*, and *QGw.cib-6A* were all contributed by CM42.

For grain number-related traits, three QTL *QFsn.cib-6A*, *QGnfs.cib-6A*, and *QSc.cib-6A* were also identified (Figures 5B, C, E). *QFsn.cib-6A* (LOD = 5.22–7.40; PVE = 7.48%–10.34%), a major and stable QTL, was detected in E5, E7, and the BLUE dataset. *QSc.cib-6A* (LOD = 4.18–7.07; PVE = 6.21%–9.86%), a stable QTL, was identified in E4, E6, and E7. *QGnfs.cib-6A* (LOD = 3.17–15.55; PVE = 4.69%–4.99%), a minor QTL, was detected in E5 and the BLUE dataset. The favorable allele of *QFsn.cib-6A*, *QGnfs.cib-6A*, and *QSc.cib-6A* was all contributed by ZKM13F10.

Meanwhile, *QPht.cib-6A*, a stable QTL, was detected in six environments (Figure 5D). It explained 6.62%–8.39% of phenotypic variance with the LOD values ranging from 4.80 to 36.31. The favorable allele of *QPht.cib-6A* was contributed by CM42.

Based on the mapping result, eight QTL, *QGns.cib-6A*, *QTgw.cib-6A*, *QGl.cib-6A*, *QGw.cib-6A*, *QFsn.cib-6A*, *QGnfs.cib-6A*, *QSc.cib-6A*, and *QPht.cib-6A*, were detected in the same interval. Temporarily, we designated this common locus as *QClu.cib-6A*.

In the QE interaction analysis, a total of 19 QTL were detected, including the nine QTL identified in the single-environment

10.3389/fpls.2023.1305547

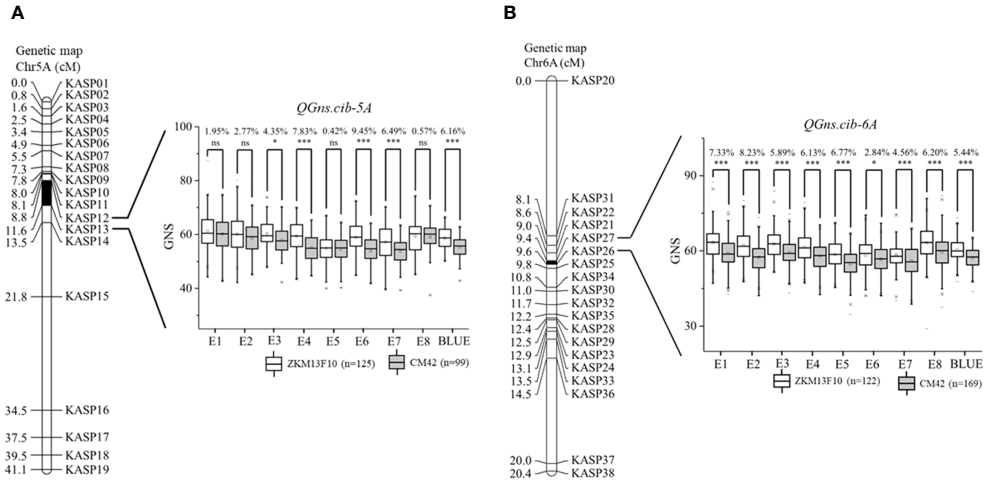


FIGURE 3
Genetic maps of the major quantitative trait loci (QTL) *QGns.cib-5A* **(A)** and *QGns.cib-6A* **(B)** and their effects on grain number per spike (GNS). The black area represents the genetic intervals of *QGns.cib-5A* and *QGns.cib-6A*. The effects of the major QTL on GNS are shown as box plots, calculated after grouping the 13CM population into two classes based on the KASP marker. ZKM13F10 and CM42 indicate the lines with and without positive alleles of *QGns.cib-5A* and *QGns.cib-6A*, respectively. ns, *, and *** represent significance at $P > 0.05$, $P < 0.05$, and $P < 0.001$.

analysis. Except for *QGns.cib-5A*, *QGw.cib-6A*, and *QPht.cib-6A*, the PVE (A) of the remaining six QTL were significantly smaller than that of PVE (AE), indicating that these QTL were not stable across environments (Supplementary Table 6). No epistatic QTL were found in this study (data not shown).

Effects of major QTL on corresponding traits

By analyzing the genotyping results of the flanking markers *KASP12* and *KASP26*, the effects of *QGns.cib-5A* and *QGns.cib-6A*

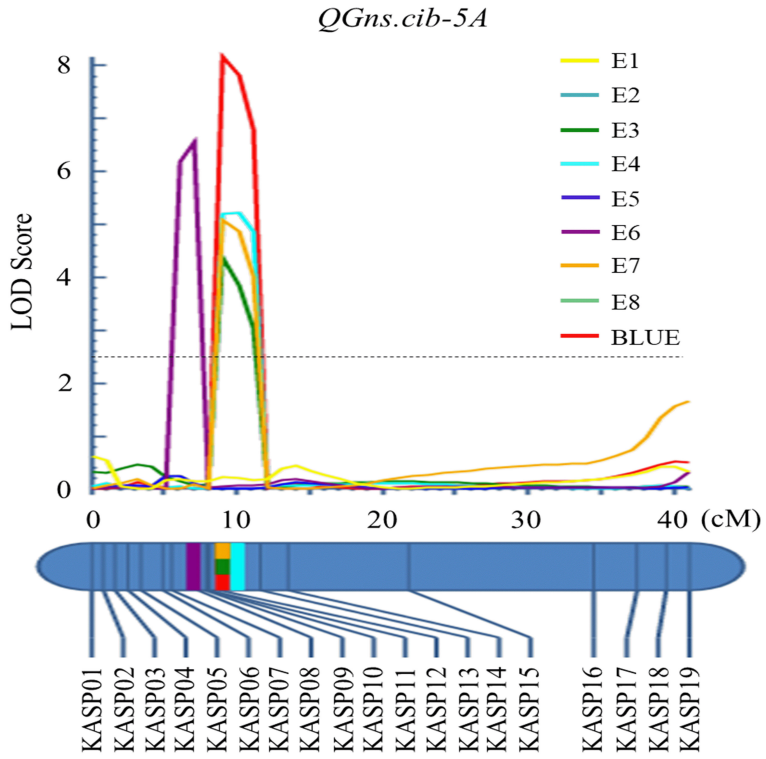


FIGURE 4
Logarithm of the odds (LOD) profile of QGns.cib-5A in the 13CM population using the inclusive composite interval mapping (ICIM) method.

TABLE 2 Quantitative trait loci (QTL) on chromosomes 5A and 6A for grain number per spike (GNS) identified across multiple environments and the BLUE dataset in the 13CM population.

QTL	Env.	Genetic interval (cM)	Flanking markers	LOD	PVE (%)	Add	Physical position (Mb)
QGns.cib-5A	E3	8.50–10.50	*KASP12–KASP13*	4.35	8.46	1.70	435.62–441.15
	E4	8.50–10.50	*KASP12–KASP13*	5.20	9.58	2.00	
	E6	8.50–10.50	*KASP07–KASP08*	6.55	12.24	2.22	
	E7	5.50–7.50	*KASP12–KASP13*	5.07	9.29	1.96	
	BLUE	8.50–10.50	*KASP12–KASP13*	8.16	14.43	1.78	
QGns.cib-6A	E5	8.50–9.50	*KASP26–KASP27*	30.80	11.47	4.05	236.95–263.29
	E7	8.50–9.50	*KASP26–KASP27*	3.82	5.44	1.40	
	E8	8.50–9.50	*KASP26–KASP27*	22.95	5.58	4.36	
	BLUE	8.50–9.50	*KASP26–KASP27*	9.04	12.38	1.53	

Env., environment; PVE, phenotypic variation explained; LOD, logarithm of the odd; Add, additive effect (positive values indicate that the alleles from ZKM13F10 increases the trait scores, and negative values indicate that the allele from CM42 increases the trait scores); BLUE, best linear unbiased estimation.

on GNS were examined. For *QGns.cib-5A*, lines with homozygous alleles from ZKM13F10 and CM42 were divided into two groups. Significant differences ($P < 0.05$ or $P < 0.001$) in GNS were observed between these groups. *QGns.cib-5A* was found to significantly increase GNS by 4.35%–9.45% across five environments (E3, E4, E6, E7, and the BLUE dataset) (Figure 3A). For *QGns.cib-6A*, significant differences ($P < 0.001$) in GNS were observed between the groups in all environments. *QGns.cib-6A* significantly increased GNS by 2.84%–8.23% (Figure 3B).

Effects of *QClu.cib-6A*, a QTL cluster, on other seven yield-related traits except GNS were assessed. For three grain size and weight traits, significant differences ($P < 0.01$ or $P < 0.001$) between groups in all or eight environments were detected and *QClu.cib-6A* significantly increased TGW, GL, and GW by 4.66%–15.92%,

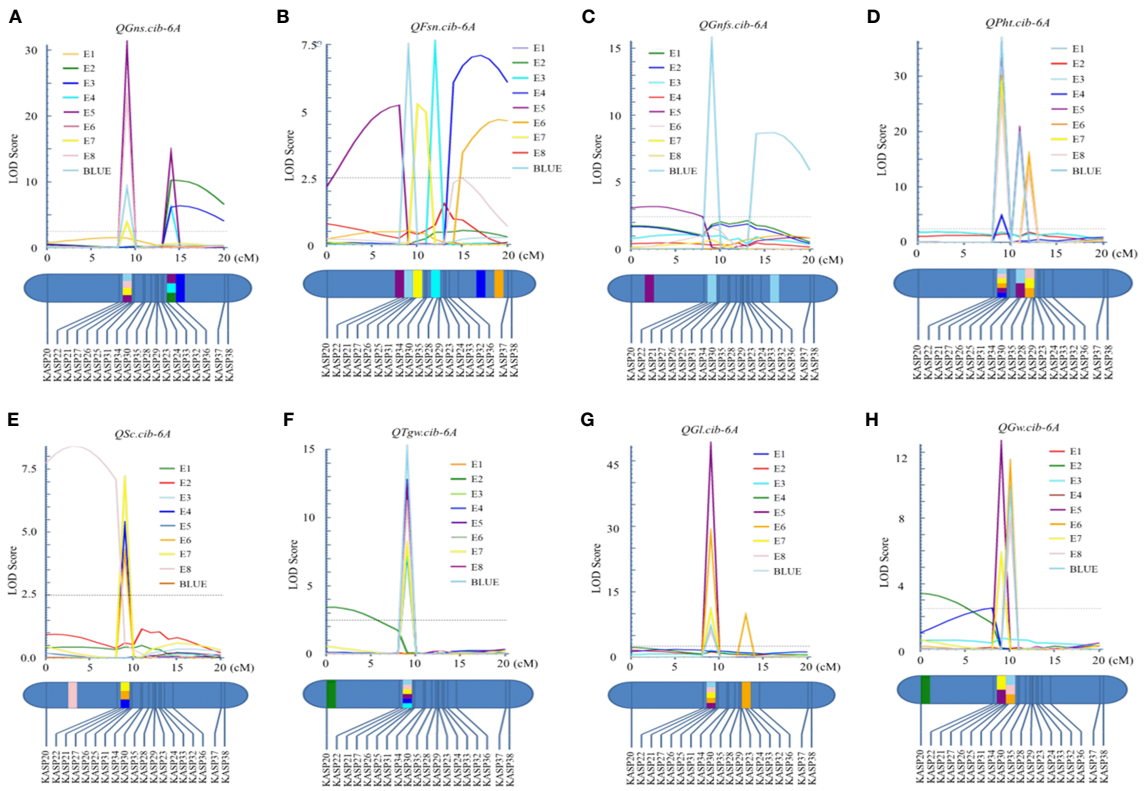


FIGURE 5
Logarithm of the odds (LOD) profile of *QGns.cib-6A* **(A)**, *QFsn.cib-6A* **(B)**, *QGnfs.cib-6A* **(C)**, *QPht.cib-6A* **(D)**, *QSc.cib-6A* **(E)**, *QTgw.cib-6A* **(F)**, *QGl.cib-6A* **(G)**, and *QGw.cib-6A* **(H)** in the 13CM population using the inclusive composite interval mapping (ICIM) method.

1.55%–4.66%, and 1.70%–6.13%, respectively (Supplementary Figures 2E-G). For three grain number-related traits, significant differences ($P < 0.05$, $P < 0.01$, or $P < 0.001$) between groups were detected in seven or six environments and *QClu.cib-6A* significantly increased FSN, GNFS, and SC by 2.72%–3.44%, 2.34%–3.76%, and 4.90%–7.12%, respectively (Supplementary Figures 2B–D). Meanwhile, significant differences ($P < 0.05$ or $P < 0.001$) on plant height were found between groups in all environments, and *QClu.cib-6A* significantly increased PHT by 3.82%–11.61% (Supplementary Figure 2A).

Effects of *QGns.cib-5A* and *QGns.cib-6A* on other yield-related traits

To detect the effects of *QGns.cib-5A* and *QGns.cib-6A* on other yield-related traits, the 13CM lines were divided into two groups based on the marker's spectra of *KASP12* and *KASP26*, respectively. For *QGns.cib-5A*, the comparative analysis between the two groups based on the BLUE dataset showed that *QGns.cib-5A* had significant effects on PTN, FSN, SL, and GL/GW ($P < 0.05$, $P < 0.01$, or $P < 0.001$) (Supplementary Figure 3). Significant differences in PTN and SL were observed between the two groups for *QGns.cib-6A* ($P < 0.001$) (Supplementary Figure 4).

Additive effect of *QGns.cib-5A* and *QGns.cib-6A*

In the present study, two QTL *QGns.cib-5A* and *QGns.cib-6A* for GNS were identified. Subsequently, the additive effects of *QGns.cib-5A* and *QGns.cib-6A* on GNS in the 13CM population were analyzed. Compared with lines with unfavorable alleles, lines with a favorable allele of GNS at *QGns.cib-5A* or *QGns.cib-6A* significantly increased GNS by 6.16% ($P < 0.001$) or 5.67% ($P < 0.001$), respectively. Compared with lines carrying unfavorable alleles, lines with both favorable alleles exhibited a significant increase in GNS by 12.85% ($P < 0.001$) (Figure 6).

Validation of *QGns.cib-5A* and *QGns.cib-6A* in different genetic backgrounds

Two populations (CZ5782 and CS104) were used to evaluate the effects of *QGns.cib-5A* and *QGns.cib-6A* in different genetic backgrounds, respectively. *KASP12* (closely linked to *QGns.cib-5A*) and *KASP26* (tightly linked to *QGns.cib-6A*) were used for genotyping. For *KASP12*, polymorphism was detected in *QGns.cib-5A* between CM104 and ZM5782. For *KASP26*, polymorphism was detected in *QGns.cib-6A* between CM104 and SH352. According to the genotyping results, the F_2 individuals from CZ5782 and CS104 were divided into three groups: individuals with a CM42 homozygous allele, individuals with a non-CM42 homozygous allele, and individuals with heterozygous allele. Significant differences ($P < 0.05$, $P < 0.01$, or $P < 0.001$) in GNS were identified between the groups with different alleles in both populations. Lines with the favorable and homozygous alleles significantly increased GNS by 6.44%–8.42% and 3.72%–8.86% in CZ5782 and CS104 populations, respectively (Figure 7).

Candidate gene analysis of *QGns.cib-5A* and *QGns.cib-6A*

After screening the physical interval of *QGns.cib-5A* (435.62 Mb–441.15 Mb) using IWGSC RefSeq v2.1, 150 prediction genes

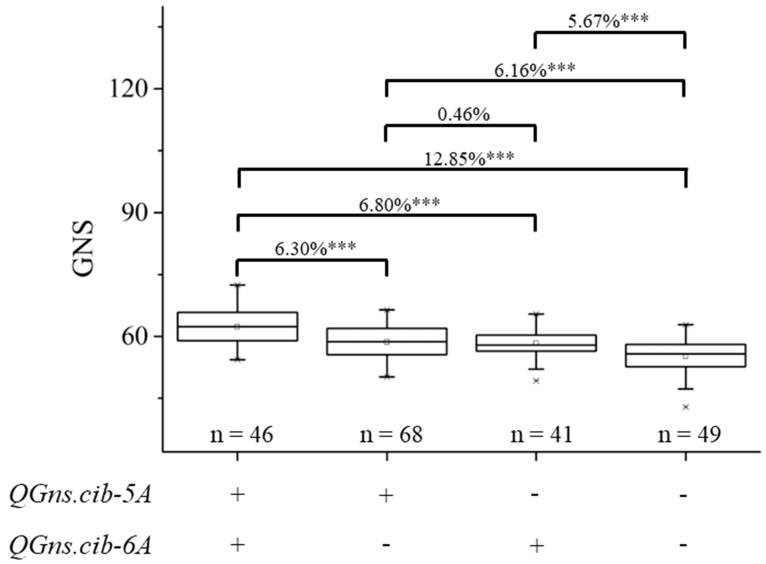


FIGURE 6
Additive effects of *QGns.cib-5A* and *QGns.cib-6A* on grain number per spike (GNS) in the 13CM population. "+" and "−" represent lines with the alleles from ZKM13F10 and CM42 of the target loci, respectively. *** represents significance at P < 0.001.

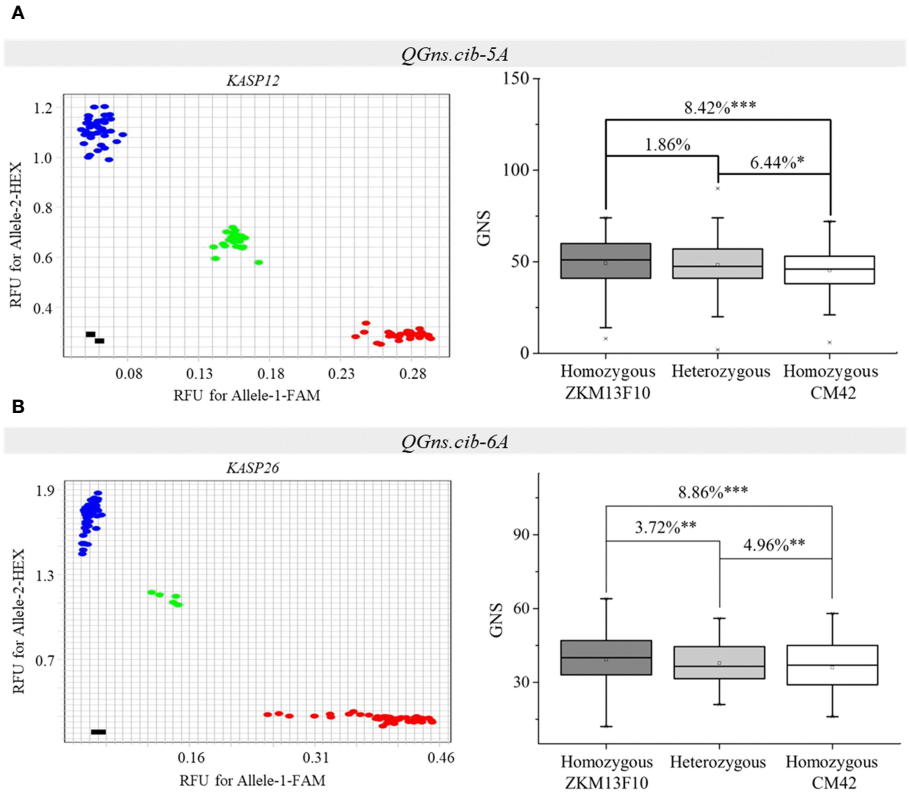


FIGURE 7
Validation of *QGns.cib-5A* **(A)** and *QGns.cib-6A* **(B)** in different genetic backgrounds. The fluorescence PCR genotyping results of the Kompetitive Allele-Specific PCR (KASP) markers *KASP12* and *KASP26* in the CZ5782 and CS352 populations, respectively. The effects of *QGns.cib-5A* and *QGns.cib-6A* on grain number per spike (GNS) in the CZ5782 and CS352 populations, respectively. *, **, and *** represent significance at $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively.

including 76 high-confidence prediction genes were obtained (Supplementary Table 7). Spatial expression patterns showed that 26 genes were highly expressed in spike, indicating that they might participate in spike development (Supplementary Figure 5). In addition, according to gene annotation, and homologous gene function in rice and/or *Arabidopsis thaliana*, *TraesCS5A03G0562600* might be related to spike development. According to the BSE-Seq data, two SNPs and an InDel were identified between the two parents of *TraesCS5A03G0562600* (Supplementary Table 8).

For *QGns.cib-6A*, 144 prediction genes (including 35 high-confidence genes) were detected in the physical interval of 236.95 Mb–263.29 Mb using IWGSC RefSeq v2.1 (Supplementary Table 7). Expression patterns suggested that 16 genes were highly expressed in spike, indicating that they might be related to spike development (Supplementary Figure 6). Furthermore, according to the gene annotation and the homologous gene function in rice and/or *Arabidopsis thaliana*, *TraesCS6A03G0487300* and *TraesCS6A03G0492700* might participate in spike development. Based on the BSE-Seq data, an InDel and one InDel were found in the upstream and exon of the *TraesCS6A03G0487300* and *TraesCS6A03G0492700*, respectively.

Haplotype analysis of *QGns.cib-5A* and *QGns.cib-6A*

According to the high-quality resequencing data of 145 Chinese wheat accessions, the haplotypes in the key regions of *QGns.cib-5A* and *QGns.cib-6A* were analyzed. Six and three haplotypes were found in *QGns.cib-5A* and *QGns.cib-6A*, respectively (Supplementary Figures 7, 8). For *QGns.cib-5A*, six KASP markers were successfully developed to differentiate the six haplotypes and used to perform the haplotype analysis in our natural population (321 wheat accessions). As expected, all six haplotypes were detected, namely, haplotype-I, -II, -III, -IV, -V, and -VI (Supplementary Figure 9A and Supplementary Table 11). Based on the association analysis result, GNS of accessions with hap-V (including ZKM13F10) was 12.27% and 2.10% higher than that of accessions with hap-VI (including CM42) in 'Cultivars' and 'Landraces', respectively (Figure 8A).

For *QGns.cib-6A*, three KASP markers were developed to distinguish the three haplotypes and haplotype analysis was carried out in 321 wheat accessions. As expected, three haplotypes (hap-I, -II, and -III) were detected (Supplementary Figure S9B; Supplementary Table 11). According to association

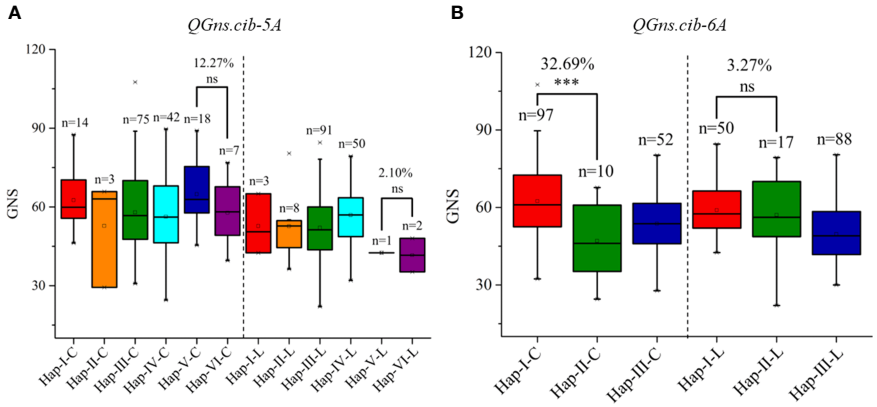


FIGURE 8
Haplotypes and their distribution frequency of *QGns.cib-5A* **(A)** and *QGns.cib-6A* **(B)** in 321 wheat accessions. 'C' and 'L' represent 'cultivars' and 'landraces', respectively; ns and *** represent significance at $P > 0.05$ and $P < 0.001$.

analysis, significant difference on GNS was detected between accessions with hap-I (including ZKM13F10) and hap-II (including CM42) in 'Cultivars'. However, a higher but non-significant difference was detected between the two haplotypes in 'Landraces' (Figure 8B).

Discussion

Comparison of the detected QTL to those reported in previous studies

In this study, two major and stable QTL, *QGns.cib-5A* and *QGns.cib-6A*, were identified on chromosomes 5A and 6A, respectively, using the BSE-Seq method and linkage analysis. To detect if they overlap with QTL reported in previous studies, we compared their physical intervals based on the CS reference genome (Table 2 and Supplementary Table 9).

For *QGns.cib-5A*, 14 QTL controlling GNS on chromosome 5A were screened in previous studies. Among them, *QGN.perg-5A* (462.01 Mb) and an unnamed QTL (tightly linked marker *Xgwm186*, 472.16 Mb) are located near the physical interval of *QGns.cib-5A* (435.62 Mb–441.15 Mb) (Liu et al., 2006; Pretini et al., 2021). *QGns.cau-5A.2* (439.67 Mb) and an un-named QTL (*Xgwm415–Xgwm304*, 107.10 Mb–664.99 Mb) were overlapped with *QGns.cib-5A* (Su et al., 2009; Guan et al., 2018). However, the unnamed QTL (*Xgwm415–Xgwm304*, 107.10 Mb–664.99 Mb) was identified in a large interval and only detected in two environments. Another unnamed QTL (tightly linked marker *Xgwm186*, 472.16 Mb) was detected in only one environment, suggesting it was unstable. *QGns.cau-5A.2* was detected in six environments and located within the physical interval of *QGns.cib-5A*, but its PVE value is less than 10%, indicating that it is a minor QTL. *QGN.perg-5A* is located near *QGns.cib-5A* and was detected in three environments with a PVE value ranging from 14.6% to 17.5%, suggesting that it is a major and stable QTL. As a result, whether *QGns.cib-5A* is a novel QTL or allelic to the reported loci remains to be revealed.

For *QGns.cib-6A*, several cloned genes and QTL around the candidate region associated with GNS have been reported in previous studies (Supplementary Tables 9, 10). *TaBT1-6A*, located near *QGns.cib-6A* and associated with grain size, weight, and grain total starch content, was identified (Wang et al., 2019). However, our remapping result showed that *TaBT1-6A* was not linked to *QGns.cib-6A*, suggesting that *TaBT1-6A* is not the candidate gene of *QGns.cib-6A*. Another gene, *TaGW2-6A*, is located within the physical interval of *QGns.cib-6A* and plays pleotropic effects on wheat agronomic traits (Su et al., 2011; Jaiswal et al., 2015). Based on an SNP site (−593 bp, A/G) in the promoter region, a KASP marker *TaGW2-6A-593* was employed (Su et al., 2011) (Supplementary Table 12). The remapping results indicated that *TaGW2-6A* was not linked to *QGns.cib-6A*, suggesting that *TaGW2-6A* is not the candidate gene of *QGns.cib-6A*. Meanwhile, no previously reported QTL for GNS overlapped with *QGns.cib-6A* (Supplementary Table 9), indicating that it may be a novel QTL.

Relationships between GNS and TGW and pleiotropic effects of *QGns.cib-5A* and *QGns.cib-6A*

Generally, a tradeoff between grain number and grain weight is usually detected, which has been a major limitation in further breeding program. In the present study, the significant and negative correlations between GNS and TGW, GL, GW, and GL/GW supports the tradeoff effect. According to statistics, approximately 90% of the identified QTL controlling GNS have a negative and pleiotropic effect on TGW (Yang et al., 2021). As a result, QTL controlling GNS with no effect on TGW is essential for breeding. In this study, *QGns.cib-6A* showed a significant and negative effect on TGW, indicating a typical tradeoff effect between GNS and TGW (Supplementary Figure 2E). On the other hand, *QGns.cib-5A* had no significant effect on TGW (Supplementary Figure 3G). This suggests that *QGns.cib-5A* can increase GNS without reducing TGW and can be utilized in breeding program.

QGns.cib-5A and *QGns.cib-6A* are the artificial selection loci during wheat improvement

During the long history of wheat domestication and selection, favorable haplotypes have been retained and enriched. However, the limited availability of genomic information has restricted access to haplotype information in wheat. Recently, more resequencing data of wheat materials have become available. For instance, the Wheat SnpHub Portal database has collected 13 resequencing datasets, encompassing 3,253 wheat accessions. This provides us the opportunity to analyze the haplotypes of a specific genomic region.

In the present study, the haplotypes for the crucial regions of *QGns.cib-5A* and *QGns.cib-6A* were analyzed using the resequencing data from 145 Chinese landmark cultivars (Hao et al., 2020). Based on the haplotype analysis of 321 wheat accessions, only one (0.65%) and two (1.29%) wheat accessions, respectively, were detected in hap-V (containing ZKM13F10) and hap-VI (containing CM42) in landraces, which suggests they were both rare haplotypes of *QGns.cib-5A*. In cultivars, the distribution frequency of the two haplotypes increased, with 18 accessions (hap-V, 11.32%) and 7 accessions (hap-VI, 4.40%), respectively. This finding indicates that both haplotypes have been artificially selected and enriched during wheat improvement. For *QGns.cib-6A*, 50 (32.26%) and 17 (10.97%) wheat accessions were found in hap-I (containing ZKM13F10) and hap-II (containing CM42) in landrace, respectively. Moreover, in cultivar, the distribution frequency of hap-I was doubled (97, 61.01%) whereas hap-II was retained (10, 6.29%). These results suggest that hap-I was enriched. Overall, both *QGns.cib-5A* and *QGns.cib-6A* appear to have been the targets of artificial selection in wheat improvement.

Potential candidate genes for *QGns.cib-5A* and *QGns.cib-6A*

Within the physical interval of *QGns.cib-5A* and *QGns.cib-6A*, 76 and 35 high-confidence prediction genes were detected in the CS reference genome, respectively (Supplementary Table 7). Through spatiotemporal expression patterns, homology analysis, function annotation, and sequence difference analysis, we predicted *TraesCS5A03G0562600* as a potential candidate gene for *QGns.cib-5A*. *TraesCS5A03G0562600* is the orthologous gene of *AUXIN RESISTANT 4* (*AXR4*) in *Arabidopsis*, and it encodes the pseudomolecule protein AUXIN RESPONSE 4 (Supplementary Table 7). In *Arabidopsis*, *AXR4* participates in biological processes of auxin polar transport (Hobbie, 2006). Auxin plays a crucial role in regulating plant growth, including the development of reproductive organs (Lampugnani et al., 2013). The BSE-Seq data revealed that two SNPs exist in the exon region of *TraesCS5A03G0562600*, which may result in functional change of this gene.

For *QGns.cib-6A*, we predicted *TraesCS6A03G0487300* and *TraesCS6A03G0492700* as potential candidate genes. *TraesCS6A03G0487300* is the orthologous gene of *SPATULA* (*SPT*) in *Arabidopsis*, encoding the pseudomolecule protein basic helix–loop–helix (bHLH) DNA-binding superfamily (Supplementary Table 7). In *Arabidopsis*, *SPT* participates in biological processes of flower development, suggesting it involves in regulating seed number (Pfannebecker et al., 2017). The BSE-Seq data revealed the presence of one SNP in the upstream region and two SNPs in the downstream regions of *TraesCS6A03G0487300*, potentially resulting in changes in expression levels (Table S8). *TraesCS6A03G0492700* is the orthologous gene of *OsUBP15* in rice and *UBIQUITIN-SPECIFIC PROTEASE 15* (*UBP15*) in *Arabidopsis*, and it encodes the pseudomolecule protein ubiquitin carboxyl-terminal hydrolase. *OsUBP15* involves in regulating the number of lateral cells in the glume and TGW in rice (Shi et al., 2019). *UBP15* participates in biological processes of cell division, flower development, fruit development, leaf development, and protein de-ubiquitination in *Arabidopsis* (Wu et al., 2022). Meanwhile, according to BSE-Seq data, two InDels were detected between the parents of *TraesCS6A03G0492700* (Supplementary Table 8). In summary, *TraesCS5A03G0562600* may be the candidate gene for *QGns.cib-5A*, whereas *TraesCS6A03G0487300* and *TraesCS6A03G0492700* may be the candidate genes for *QGns.cib-6A*, and their further investigation through map-based cloning would be valuable.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

CJ: Data curation, Formal Analysis, Investigation, Software, Validation, Writing – original draft. ZX: Data curation, Resources, Writing – review & editing. XF: Methodology, Software, Writing – review & editing. QZ: Data curation, Resources, Writing – review & editing. GJ: Data curation, Resources, Writing – review & editing. SL: Data curation, Software, Writing – review & editing. YW: Software, Visualization, Writing – review & editing. FM: Software, Writing – review & editing. YZ: Project administration, Supervision, Writing – review & editing. TW: Funding acquisition, Project administration, Writing – review & editing. BF: Funding acquisition, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Sichuan Science and Technology Program, China (2022ZDZX0016), the West Light Foundation of the Chinese Academy of Sciences (2022XBZG_XBQNXZ_A_001), and the Major Science and Technology Achievement Transformation of Central Universities and Institutes in Sichuan Projects (2022ZHCG0131).

Acknowledgments

We express our gratitude to the Triticeae Multi-omics Center (http://202.194.139.32/) for providing us with an integrated platform of tools and genomic data, which greatly facilitated our research. We also acknowledge the Wheat-SnpHub-Portal (http://wheat.cau.edu.cn/Wheat_SnpHub_Portal/) for providing the genomic variation datasets of wheat, and Bioacme Biotechnology Co., Ltd. (Wuhan, China, http://www.whbioacme.com) for their assistance with the BSE-Seq analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1305547/full#supplementary-material

References

Azadi, A., Mardi, M., Hervan, E. M., Mohammadi, S. A., Moradi, F., Tabatabaee, M. T., et al. (2014). QTL Mapping of yield and yield components under normal and salt-stress conditions in bread wheat (*Triticum aestivum* L.). *Plant Mol. Biol. Rep.* 33, 102–120. doi: 10.1007/s11105-014-0726-0

Backhaus, A. E., Lister, A., Tomkins, M., Adamski, N. M., Simmonds, J., Macaulay, I., et al. (2022). High expression of the MADS-box gene *VRT2* increases the number of rudimentary basal spikelets in wheat. *Plant Physiol.* 189, 1536e1552. doi: 10.1093/plphys/kiac156

Blanco, A., Mangini, G., and Giancaspro, A. (2012). Relationships between grain protein content and grain yield components through quantitative trait locus analyses in a recombinant inbred line population derived from two elite durum wheat cultivars. *Mol. Breed.* 30, 79–92. doi: 10.1007/s11032-011-9600-z

Boden, S. A., Cavanagh, C., Cullis, B. R., Ramm, K., Greenwood, J., Finnegan, E. J., et al. (2015). *Ppd-1* is a key regulator of inflorescence architecture and paired spikelet development in wheat. *Nat. Plants* 1, 14016. doi: 10.1038/nplants.2014.16

Börner, A., Schumann, E., Fürste, A., Cöster, H., Leithold, B., Röder, M., et al. (2002). Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 105, 921–936. doi: 10.1007/s00122-002-0994-1

Chen, C. J., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y. H., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009

Chen, T. T., Chen, X., Zhang, S. S., Zhu, J. W., Tang, B. X., Wang, A. K., et al. (2021). The genome sequence archive family: toward explosive data growth and diverse data types. *Genom. Proteom. Bioinf.* 19, 578–583. doi: 10.1016/j.gpb.2021.08.001

Chen, A., and Dubcovsky, J. (2012). Wheat TILLING mutants show that the vernalization gene *VRN1* down-regulates the flowering repressor *VRN2* in leaves but is not essential for flowering. *PloS Genet.* 8, e1003134. doi: 10.1371/journal.pgen.1003134

CNCB-NGDC Members and Partners (2022). Database resources of the national genomics data center, China national center for bioinformation in 2022. *Nucleic Acids Res.* 50, D27–D38. doi: 10.1093/nar/gkab951

Cuthbert, J. L., Somers, D. J., and Brûlé-Babel, A. L. (2008). Molecular mapping of quantitative trait loci for yield and yield components in spring wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 117, 595–608. doi: 10.1007/s00122-008-0804-5

Debernardi, J. M., Greenwood, J. R., Jean, F. E., Jernstedt, J., and Dubcovsky, J. (2020). *APETALA 2*-like genes *AP2L2* and *Q* specify lemma identity and axillary floral meristem development in wheat. *Plant J.* 101, 171e187. doi: 10.1111/tpj.14528

Debernardi, J. M., Lin, H., Chuck, G., Faris, J. D., and Dubcovsky, J. (2017). microRNA172 plays a crucial role in wheat spike morphogenesis and grain threshability. *Development* 144, 1966e1975. doi: 10.1242/dev.146399

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., and Yanofsky, M. F. (2004). The *SEP4* gene of *Arabidopsis thaliana* functions in floral organ and meristem identity. *Curr. Biol.* 14, 1935e1940. doi: 10.1016/j.cub.2004.10.028

Dixon, L. E., Greenwood, J. R., Bencivenga, S., Zhang, P., Cockram, J., Mellers, G., et al. (2018). *TEOSINTE BRANCHED1* regulates inflorescence architecture and development in bread wheat (*Triticum aestivum*). *Plant Cell* 30, 563e581. doi: 10.1105/tpc.17.00961

Du, D. J., Zhang, D. X., Yuan, J., Feng, M., Li, Z. J., Wang, Z. H., et al. (2021). *FRIZZY PANICLE* defines a regulatory hub for simultaneously controlling spikelet formation and awn elongation in bread wheat. *New Phytol.* 231, 814e833. doi: 10.1111/nph.17388

Gao, C. X. (2021). Genome engineering for crop improvement and future agriculture. *Cell* 184, 1621–1635. doi: 10.1016/j.cell.2021.01.005

Gao, F. M., Wen, W. E., Liu, J. D., Rasheed, A., Yin, G. H., Xia, X. C., et al. (2015). Genome-wide linkage mapping of QTL for yield components, plant height and yield-related physiological traits in the Chinese wheat cross Zhou 8425B/Chinese Spring. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01099

Guan, P. F., Lu, L. H., Jia, L. J., Kabir, M. R., Zhang, J. B., Lan, T. Y., et al. (2018). Global QTL analysis identifies genomic regions on chromosomes 4A and 4B harboring stable loci for yield-related traits across different environments in wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00529

Hao, C. Y., Jiao, C. Z., Hou, J., Li, T., Liu, H. X., Wang, Y. Q., et al. (2020). Resequencing of 145 landmark cultivars reveals asymmetric sub-genome selection and strong founder genotype effects on wheat breeding in China. *Mol. Plant* 13, 1733–1751. doi: 10.1016/j.molp.2020.09.001

Hobbie, L. J. (2006). Auxin and cell polarity: the emergence of *AXR4*. *Trends Plant Sci.* 11, 517–518. doi: 10.1016/j.tplants.2006.09.003

Hu, W. J., Gao, D. R., Liao, S., Cheng, S. H., Jia, J. Z., and Xu, W. G. (2023). Identification of a pleiotropic QTL cluster for Fusarium head blight resistance, spikelet compactness, grain number per spike and thousand-grain weight in common wheat. *Crop J.* 11, 672–677. doi: 10.1016/j.cj.2022.09.007

Hu, J. M., Wang, X. Q., Zhang, G. X., Jiang, P., Chen, W. Y., Hao, Y. C., et al. (2020). QTL mapping for yield-related traits in wheat based on four RIL populations. *Theor. Appl. Genet.* 133, 917–933. doi: 10.1007/s00122-019-03515-w

Huang, X. Q., Kempf, H., Ganal, M. W., and Röder, M. S. (2004). Advanced backcross QTL analysis in progenies derived from a cross between a German elite winter wheat variety and a synthetic wheat (*Triticum aestivum*L.). *Theor. Appl. Genet.* 109, 933–943. doi: 10.1007/s00122-004-1708-7

Jaiswal, V., Gahlaut, V., Mathur, S., Agawai, P., Khandelwal, M. K., Khurana, J. P., et al. (2015). Identification of novel SNP in promoter sequence of *TaGW2-6A* associated with grain weight and other agronomic traits in wheat (*Triticum aestivum* L.). *PloS One* 10, e0129400. doi: 10.1371/journal.pone.0129400

Jiang, C., Xu, Z. B., Fan, X. L., Zhou, Q., Ji, G. S., Chen, L. E., et al. (2023). Identification and validation of quantitative trait loci for fertile spikelet number per spike and grain number per fertile spikelet in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 136, 69. doi: 10.1007/s00122-023-04297-y

Ji, G. S., Xu, Z. B., Fan, X. L., Zhou, Q., Chen, L. E., Yu, Q., et al. (2023). Identification and validation of major QTL for grain size and weight in bread wheat (*Triticum aestivum* L.). *Crop J.* 11, 564–572. doi: 10.1016/j.cj.2022.06.014

Ji, G. S., Xu, Z. B., Fan, X. L., Zhou, Q., Yu, Q., Liu, X. F., et al. (2021). Identification of a major and stable QTL on chromosome 5A confers spike length in wheat (*Triticum aestivum* L.). *Mol. Breed.* 41, 56. doi: 10.1007/s11032-021-01249-6

Kong, X. C., Wang, F., Geng, S. F., Guan, J. T., Tao, S., Jia, M. L., et al. (2021). The wheat *AGL6*-like MADS-box gene is a master regulator for floral organ identity and a target for spikelet meristem development manipulation. *Plant Biotechnol. J.* 20, 75–88. doi: 10.1111/pbi.13696

Kumar, N., Kulwal, P. L., Balyan, H. S., and Gupta, P. K. (2007). QTL mapping for yield and yield contributing traits in two mapping populations of bread wheat. *Mol. Breed.* 19, 163–177. doi: 10.1007/s11032-006-9056-8

Lampugnani, E. R., Kilinc, A., and Smyth, D. R. (2013). Auxin controls petal initiation in *Arabidopsis*. *Development* 140, 185–194. doi: 10.1242/dev.084582

Li, C. X., and Dubcovsky, J. (2008). Wheat FT protein regulates *VRN1* transcription through interactions with FDL2. *Plant J.* 55, 543e554. doi: 10.1111/j.1365-313X.2008.03526.x

Li, A. L., Hao, C. Y., Wang, Z. Y., Geng, S. F., Jia, M. L., Wang, F., et al. (2022). Wheat breeding history reveals synergistic selection of pleiotropic genomic sites for plant architecture and grain yield. *Mol. Plant* 15, 504–519. doi: 10.1016/j.molp.2022.01.004

Li, Y. P., Li, L., Zhao, M. C., Guo, L., Guo, X. X., Zhao, D., et al. (2021). Wheat *FRIZZY PANICLE* activates *VERNALIZATION1-A* and *HOMEOBOX4-A* to regulate spike development in wheat. *Plant Biotechnol. J.* 19, 1141e1154. doi: 10.1111/pbi.13535

Li, C. X., Lin, H. Q., Chen, A., Lau, M., Jernstedt, J., and Dubcovsky, J. (2019). Wheat *VRN1*, *FUL2* and *FUL3* play critical and redundant roles in spikelet development and spike determinacy. *Development* 146, dev175398. doi: 10.1242/dev.175398

Li, L., Shi, F., Wang, Y. Q., Yu, X. F., Zhi, J. J., Guan, Y. B., et al. (2020). TaSPL13 regulates inflorescence architecture and development in transgenic wheat (*Triticum aestivum* L.). *Plant Sci.* 296, 110516. doi: 10.1016/j.plantsci.2020.110516

Liu, J., Chen, Z. Y., Wang, Z. H., Zhang, Z. H., Xie, X. M., Wang, Z. H., et al. (2021). Ectopic expression of *VRT-A2* underlies the origin of *Triticum polonicum* and *Triticum petropavlovskyi* with long outer glumes and grains. *Mol. Plant* 14, 1472e1488. doi: 10.1016/j.molp.2021.05.021

Liu, J., Cheng, X. L., Liu, P., and Sun, J. Q. (2017). miR156-Targeted SBP-box transcription factors interact with DWARF53 to regulate *TEOSINTE BRANCHED1* and *BARREN STALK1* expression in bread wheat. *Plant Physiol.* 174, 1931e1948. doi: 10.1104/pp.17.00445

Liu, C. Y., Sukumaran, S., Claverie, E., Sansaloni, C., Dreisigacker, S., and Reynolds, M. (2019). Genetic dissection of heat and drought stress QTLs in phenology-controlled synthetic-derived recombinant inbred lines in spring wheat. *Mol. Breed.* 39, 34. doi: 10.1007/s11032-019-0938-y

Liu, S. B., Zhou, R. H., Dong, Y. C., Li, P., and Jia, J. Z. (2006). Development, utilization of introgression lines using a synthetic wheat as donor. *Theor. Appl. Genet.* 112, 1360–1373. doi: 10.1007/s00122-006-0238-x

Luo, X. M., Yang, Y. M., Lin, X. L., and Xiao, J. (2023). Deciphering spike architecture formation towards yield improvement in wheat. *J. Genet. Genomics*. 50, 835–845. doi: 10.1016/j.jgg.2023.02.015

McIntyre, C. L., Mathews, K. L., and Rattey, A. (2010). Molecular detection of genomic regions associated with grain yield and yield-related components in an elite bread wheat cross evaluated under irrigated and rainfed conditions. *Theor. Appl. Genet.* 120, 527–541. doi: 10.1007/s00122-009-1173-4

Meng, L., Li, H. H., Zhang, L. Y., and Wang, J. K. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001

Mizuno, N., Ishikawa, G., Kojima, H., Tougou, M., Kiribuchi-Otobe, C., Fujita, M., et al. (2021). Genetic mechanisms determining grain number distribution along the spike and their effect on yield components in wheat. *Mol. Breed.* 41, 62. doi: 10.1007/s11032-021-01255-8

Pelaz, S., Ditta, G. S., Baumann, E., Wisman, E., and Yanofsky, M. F. (2000). B and C floral organ identity functions require *SEPALLATA* MADS-box genes. *Nature* 405, 200e203. doi: 10.1038/35012103

Peng, J. H., Ronin, Y., Fahima, T., Röder, M. S., Li, Y. C., Nevo, E., et al. (2003). Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2489–2494. doi: 10.1073/pnas.252763199

Pfannebecker, K. C., Lange, M., Rupp, O., and Becker, A. (2017). Seed plant-specific gene lineages involved in carpel development. *Mol. Biol. Evol.* 34, 925–942. doi: 10.1093/molbev/msw297

Poursarebani, N., Seidensticker, T., Koppolu, R., Trautewig, C., Gawroński, P., Bini, F., et al. (2015). The genetic basis of composite spike form in barley and 'Miracle-Wheat'. *Genetics* 201, 155e165. doi: 10.1534/genetics.115.176628

Pretini, N., Vanzetti, L. S., Terrile, I. I., Donaire, G., and González, F. G. (2021). Mapping QTL for spike fertility and related traits in two doubled haploid wheat (*Triticum aestivum* L.) populations. *BMC Plant Biol.* 21, 353. doi: 10.1186/s12870-021-03061-y

Qiao, L., Li, H. L., Wang, J., Zhao, J., Zheng, X., Wu, B., et al. (2022). Analysis of genetic regions related to field grain number per spike from Chinese wheat founder parent Linfen 5064. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.808136

Quarrie, S. A., Steed, A., and Calestani, C. (2005). A high-density genetic map of hexaploid wheat (*Triticum aestivum* L.) from the cross Chinese Spring × SQ1 and its use to compare QTLs for grain yield across a range of environments. *Theor. Appl. Genet.* 110, 865–880. doi: 10.1007/s00122-004-1902-7

Ray, D. K., Mueller, N. D., West, P. C., and Foley, J. A. (2013). Yield trends are insufficient to double global crop production by 2050. *PloS One* 8, e66428. doi: 10.1371/journal.pone.0066428

Roncallo, P. F., Alliraju, P. C., Cervigni, G. L., and Echenique, V. C. (2017). QTL mapping and analysis of epistatic interactions for grain yield and yield-related traits in *Triticum turgidum* L. var. *durum*. *Euphytica* 213, 277. doi: 10.1007/s10681-017-2058-2

Rustgi, S., Shafqat, M. N., Kumar, N., Baenziger, P. S., Ali, M. L., Dweikat, I., et al. (2013). Genetic dissection of yield and its component traits using high-density composite map of wheat chromosome 3A: bridging gaps between QTLs and underlying genes. *PloS One* 8, e70526. doi: 10.1371/journal.pone.0070526

Shaw, L. M., Lyu, B., Turner, R., Li, C., Chen, F., Han, X., et al. (2019). *FLOWERING LOCUS T2* regulates spike development and fertility in temperate cereals. *J. Exp. Bot.* 70, 193e204. doi: 10.1093/jxb/ery350

Shi, C. L., Ren, Y. L., Liu, L. L., Wang, F., Zhang, H., Tian, P., et al. (2019). *Ubiquitin Specific Protease 15* has an important role in regulating grain width and size in rice. *Plant Physiol.* 180, 381–391. doi: 10.1104/pp.19.00065

Su, Z., Hao, C., Wang, L., Dong, Y., and Zhang, X. (2011). Identification and development of a functional marker of *TaGW2* associated with grain weight in bread wheat (*Triticum aestivum* L). *Theor. Appl. Genet.* 122, 211–223. doi: 10.1007/s00122-010-1437-z

Su, J. Y., Zheng, Q., Li, H. W., Li, B., Jing, R. L., Tong, Y. P., et al. (2009). Detection of QTLs for phosphorus use efficiency in relation to agronomic performance of wheat grown under phosphorus sufficient and limited conditions. *Plant Sci.* 176, 824–836. doi: 10.1016/j.plantsci.2009.03.006

Wang, R. X., Hai, L., Zhang, X. Y., You, G. X., Yan, C. S., and Xiao, S. H. (2009). QTL mapping for grain filling rate and yield-related traits in RILs of the Chinese winter wheat population Heshangmai × Yu8679. *Theor. Appl. Genet.* 118, 313–325. doi: 10.1007/s00122-008-0901-5

Wang, Y., Hou, J., Liu, H., Li, T., Wang, K., Hao, C. Y., et al. (2019). *TaBT1* affecting starch synthesis and thousand kernel weight underwent strong selection during wheat improvement. *J. Exp. Bot.* 70, 1497–1511. doi: 10.1093/jxb/erz032

Wang, J. S., Liu, W. H., Wang, H., Li, L., Wu, J., Yang, X., et al. (2011). QTL mapping of yield-related traits in the wheat germplasm 3228. *Euphytica* 177, 277–292. doi: 10.1007/s10681-010-0267-z

Wu, X. D., Cai, X. B., Zhang, B. W., Wu, S., Wang, R., Li, N., et al. (2022). ERECTA regulates seed size independently of its intracellular domain via MAPK-DA1-UBP15 signaling. *Plant Cell* 34, 3773–3789. doi: 10.1093/plcell/koac194

Yan, L. L., Fu, D. L., Li, C. X., Blechl, A., Tranquilli, G., Bonafede, M., et al. (2006). The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. Proc. *Natl. Acad. Sci. U. S. A.* 103, 19581e19586. doi: 10.1073/pnas.0607142103

Yan, L. L., Helguera, M., Kato, K., Fukuyama, S., Sherman, J., and Dubcovsky, J. (2004). Allelic variation at the *VRN-1* promoter region in polyploid wheat. *Theor. Appl. Genet.* 109, 1677e1686. doi: 10.1007/s00122-004-1796-4

Yang, Y., Amo, A., Wei, D., Chai, Y., Zhang, J., Qiao, P., et al. (2021). Large-scale integration of meta-QTL and genome-wide association study discovers the genomic regions and candidate genes for yield and yield-related traits in bread wheat. *Theor. Appl. Genet.* 134, 3083–3109. doi: 10.1007/s00122-021-03881-4

Yu, Q., Feng, B., Xu, Z. B., Fan, X. L., Zhou, Q., Ji, G. S., et al. (2022). Genetic dissection of three major quantitative trait loci for spike compactness and length in bread wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.882655

Zhang, H., Chen, J. S., Li, R. Y., Deng, Z., Zhang, K., Liu, B., et al. (2016). Conditional QTL mapping of three yield components in common wheat (*Triticum aestivum* L.). *Crop J.* 4, 220–228. doi: 10.1016/j.cj.2016.01.007

Zhong, J. S., van Esse, G. W., Bi, X. J., Lan, T., Walla, A., Sang, Q., et al. (2021). *INTERMEDIUM-M* encodes an *HvAP2L-H5* ortholog and is required for inflorescence indeterminacy and spikelet determinacy in barley. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2011779118. doi: 10.1073/pnas.2011779118

Zhu, T. T., Wang, L., Rimbert, H., Rodriguez, J. C., Deal, K. R., De Oliveira, R., et al. (2021). Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. *Plant J.* 107, 303–314. doi: 10.1111/tpj.15289

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/fpls.2023.1304388

OPEN ACCESS

EDITED BY
Muhammad Kashif Riaz Khan,
Nuclear Institute for Agriculture and Biology, Pakistan

REVIEWED BY
Hongjian Zheng,
Shanghai Academy of Agricultural Sciences, China
Jindong Liu,
Chinese Academy of Agricultural Sciences, China
Qiaojun Lou,
Shanghai Agrobiological Gene Center, China
Kai Chen,
Chinese Academy of Agricultural Sciences, China
Naser Farrokhi,
Shahid Beheshti University, Iran

*CORRESPONDENCE
Rakesh Singh
rakesh.singh2@icar.gov.in

RECEIVED 29 September 2023
ACCEPTED 22 December 2023
PUBLISHED 11 January 2024

CITATION
Sachdeva S, Singh R, Maurya A, Singh VK, Singh UM, Kumar A and Singh GP (2024) Multi-model genome-wide association studies for appearance quality in rice. *Front. Plant Sci.* 14:1304388. doi: 10.3389/fpls.2023.1304388

Multi-model genome-wide association studies for appearance quality in rice

Supriya Sachdeva[1], Rakesh Singh[1]*, Avantika Maurya[1], Vikas Kumar Singh[2], Uma Maheshwar Singh[3], Arvind Kumar[4] and Gyanendra Pratap Singh[5]

[1]Division of Genomic Resources, ICAR-National Bureau of Plant Genetic Resources (NBPGR), New Delhi, India, [2]International Rice Research Institute, South Asia Hub, International Crop Reseach Institute for Semi Arid Tropics (ICRISAT), Hyderabad, India, [3]International Rice Research Institute, South Asia Regional Centre (ISARC), Varanasi, India, [4]International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Telangana, India, [5]Indian Council of Agricultural Research (ICAR)-National Bureau of Plant Genetic Resources, New Delhi, India

Improving the quality of the appearance of rice is critical to meet market acceptance. Mining putative quality-related genes has been geared towards the development of effective breeding approaches for rice. In the present study, two SL-GWAS (CMLM and MLM) and three ML-GWAS (FASTmrEMMA, mrMLM, and FASTmrMLM) genome-wide association studies were conducted in a subset of 3K-RGP consisting of 198 rice accessions with 553,831 SNP markers. A total of 594 SNP markers were identified using the mixed linear model method for grain quality traits. Additionally, 70 quantitative trait nucleotides (QTNs) detected by the ML-GWAS models were strongly associated with grain aroma (AR), head rice recovery (HRR, %), and percentage of grains with chalkiness (PGC, %). Finally, 39 QTNs were identified using single- and multi-locus GWAS methods. Among the 39 reliable QTNs, 20 novel QTNs were identified for the above-mentioned three quality-related traits. Based on annotation and previous studies, four functional candidate genes (*LOC_Os01g66110*, *LOC_Os01g66140*, *LOC_Os07g44910*, and *LOC_Os02g14120*) were found to influence AR, HRR (%), and PGC (%), which could be utilized in rice breeding to improve grain quality traits.

KEYWORDS

rice, grain quality, QTNs, candidate genes, GWAS

Introduction

Cultivated rice (*Oryza sativa* L.) is an important source of calories for more than half of the global population. With improved living standards and increasing awareness among people worldwide, there is a growing demand for the consumption of superior quality healthier rice varieties (Bao, 2014; Adjah et al., 2020; Selvaraj et al., 2021; Hori and Sun,

2022). Therefore, high-quality rice has become a paramount consideration for rice breeders, consumers, and producers (Qiu et al., 2021). The crucial determinants of rice grain quality include appearance, milling, nutritional composition, aroma, and cooking properties. Recently, more efforts have been made to breed rice varieties with desirable traits in terms of higher head rice recovery (HRR, %), and lower percentage of chalky grains (PGC, %) by discovering key haplotype variations, thereby harnessing allelic diversity in the germplasm (Selvaraj et al., 2021). Currently, molecular advances and genome sequencing platforms with lower costs have aided in cloning and functionally dissecting a series of genetic factors/quantitative trait loci (QTLs) in rice (Varshney et al., 2014; Abbai et al., 2019). Genetic studies have shown that multiple factors control each quality trait reflecting the intricate nature of the rice quality traits (Li et al., 2022). The genes affecting these physicochemical characteristics are related to starch biosynthesis, the metabolism of seed storage proteins (SSPs), and specific nutraceutical compounds (Biselli et al., 2015). Grain chalkiness, for example, is associated with many genes such as *Flo2*, *Chalk5*, *GIF2*, *LTPs*, *GBSS I*, *OsPUL*, *OsBT1*, *OsBE1*, and *SSIIa* (Li et al., 2014a; Wang et al., 2018), and several QTLs have been detected and widely distributed across the rice genome (Zhang H. et al., 2019; Hori et al., 2021), two of which have been fine-mapped by association and linkage mapping, such as qPGWC-7 (Zhou et al., 2009), qPGWC-8 (Guo et al., 2011; Zhao et al., 2016), and one QTL cluster mapped on chromosome 4 by single and joint mapping studies between the markers id4007289 and RM252. Loss of function mutations and genic interactions between the alleles of well-known genes responsible for biosynthesis of starch, *viz.*, *GBSSI*, *SS2a*, *SS3a*, *SS4b*, *BE2b*, and, *ISA1* gene have been shown to increase the amount of resistant starch in rice, which is believed to be crucial for improving human health (Zhang C. et al., 2019; Fujita et al., 2022; Miura et al., 2022).

Regarding the percentage of rice recovery determining rice grain quality, approximately 34 genes/QTLs have been documented in all rice chromosomes, which are largely influenced by the environment (Bao, 2014). A common QTL for grain size and head rice recovery was also detected on chromosome 3, suggesting a relationship between these two traits at the genetic level (Tan et al., 2001). An increase in grain yield has been reported in near-isogenic lines (NILs) introduced with the null allele of rice chalkiness gene *PDIL1-1*, explaining significant differences in phenotype between the genetic makeup of the rice cultivars; however, there was an increase in grain chalkiness (Hori and Sun, 2022). The appearance and rice grain quality are closely related to its rice grain size (Xie et al., 2013; Bao, 2019). Interestingly, pleiotropic effects have been reported in 25 cloned QTLs identified for multiple grain size-controlled traits, namely rice yield, appearance, and grain quality (Wang et al., 2018). Furthermore, the *gw2* WY3 allele had positive effects on grain yield, but reduced grain quality by increasing PGC (%) and decreasing HRR (%) (Song et al., 2007).

Fragrant rice is a special group with a distinct aroma, flavor, and medicinal, antioxidant, and stress-resistance properties. To date, more than 200 aroma compounds have been documented in fragrant rice (Champagne, 2008) and 2-acetly-1-pyrroline (2-AP) has been recognized as the most prominent compound contributing to aroma production in rice (Poonlaphdecha et al., 2016; Wakte et al., 2017) which is under the control of a recessive gene *Badh2*. RNA Seq studies have shown that the expression of heavy metal transporters in response to zinc at the transcriptional and post-transcriptional levels, and their epigenetic modifications, regulate the biosynthesis of 2-AP in aromatic rice varieties (Imran et al., 2022). The haplotype diversity of the *Badh2* gene was investigated in 22 fragrant landraces from Thailand, identifying four new haplotypes (H1, H2, H3, and H4). These *badh2* alleles may serve as functional markers, and landraces with a favorable haplotype (H1) could be employed as genetic resources in rice breeding programs (Chan-In et al., 2020). Several other genes affecting seed development and quality traits have been characterized, such as *GW2* (Song et al., 2007), *GS3* (Sun et al., 2018), *GS2* (Hu et al., 2015), *GS5* (Xu et al., 2015), *GS9* (Zhao et al., 2018), *GW5* (Duan et al., 2017), *GLW7* (Si et al., 2016), and *OsMAPK6* (Liu et al., 2015). Therefore, understanding the molecular basis of these traits is a prerequisite for identifying novel alleles and donors related to high grain quality, which could considerably improve rice breeding efficiency (Yano et al., 2016; Wang et al., 2017; Abbai et al., 2019; Misra et al., 2019; Verma et al., 2021; Zhong et al., 2021). These newly recognized superior versions of quality genes might then be taken together through the rapid and undoubtedly proved concept of 'haplotype introgression' (Bevan et al., 2017). Nevertheless, the lack of information regarding the superior haplotype combinations of several key grain quality genes has been one of the major bottlenecks, and the 3000-rice genome project (3K-RGP) offers enormous potential for harnessing the haplotype diversity of grain quality genes in rice (Li et al., 2014b).

Genome-wide association studies (GWAS) have become popular for the genetic dissection of complex traits into QTL/candidate genes that might be deployed in precision breeding programs aimed at crop improvement (Lipka et al., 2015, Tibbs et al., 2021). It is considered more efficient than bi-parental mapping approaches considering the naturally occurring genetic diversity, high-density genetic markers, and fewer linkage disequilibrium to identify candidate genes (Alqudah et al., 2020). Statistical methods with varying degrees of reliability substantially influence the significant MTAs determined by GWAS (Gawenda et al., 2015; Visscher et al., 2017; Wen et al., 2018). The commonly used single-locus mixed model independently scans each SNP marker for association with a phenotypic trait (Waugh et al., 2014; Gupta et al., 2019). However, this model lacks accuracy in estimating the SNP effects and identifies false negatives if the desired trait is governed by many genes at different loci (Wang et al., 2016), which is a common scenario in most quantitative traits or in case it requires a Bonferroni correction (Wen et al., 2018). It has also been proposed that single-locus models fail to detect the epistatic interactions that may exist between the closely linked genes (Gawenda et al., 2015) and are less suitable for harnessing the haplotype diversity of genes of interest that exist in the germplasm (Lu et al., 2011; Contreras-Soto et al., 2017; N'Diaye et al., 2017). To overcome the shortcomings of single-locus models, multi-locus models such as multi-locus random SNP-effect MLM (mrMLM) (Wang et al., 2016); multi-locus mixed model (MLMM) (Segura

et al., 2012), iteractive modified sure-independence screening expectation maximization Baysian least absolute shrinkage and selection operator (ISIS EM-BLASSO) (Tamba et al., 2017), FASTmrMLM (multi-locus random SNP-effect) (Tamba and Zhang, 2018), FASTmrEMMA (fast multi-locus random-SNP-effect efficient mixed model analysis) (Wang et al., 2016), polygenic-background-control-based least angle empirical Bayes (pLARmEB) (Zhang et al., 2017), and integration of Kruskal–Wallis test with empirical Bayes (pKWmEB) (Ren et al., 2018) were developed that test multiple SNP markers simultaneously to capture the molecular basis underlying different complex traits in different crop species (Wang et al., 2016) by overcoming the strong population structure and high linkage disequilibrium between the markers. In this investigation, we performed a GWAS and conducted a candidate gene-based association study in a set of 3K-RGP panels, analyzed the haplotype diversity of candidate genes, and evaluated the performance of different haplotypes associated with grain aroma, head rice recovery (HRR, %), and percentage of grains with chalkiness (PGC, %) to accelerate the design of next-generation quality-rich rice varieties by incorporating superior haplotypes for use in future rice improvement programs.

Materials and methods

Plant materials and phenotyping

A subset panel of 3K re-sequenced genomes (https://doi.org/10.1186/2047-217X-3-7) was obtained from the IRRI South Asia Regional Center, NSRTC Campus, Varanasi, Uttar Pradesh, India. The 196 rice accessions used in our investigation were collected from 89 countries belonging to four major populations: *Xian(indica)* (171), *aus/boro* (22), *tropical Geng (japonica)* (3), intermediate type (2), and two semi-dwarf varieties Pusa Basmati 1121 and PB-1 (Supplementary Table 1). The 198 accessions were planted in randomized plots in the field at the ICAR-Indian Agricultural Research Institute (IARI), New Delhi, India with four replications within Kharif 2020 and Kharif 2021. The uniform growth of seedlings was confirmed by germinating seeds on a raised seedbed, and 21 days old plantlets were transplanted. Each accession was sown in two rows, with each row consisting of 10 plants at a distance of 20 cm × 15 cm within and between the two rows. Standard practices were followed for field management. At maturity, paddy seeds from each plot were collected in bulk and dried in hot air ovens. Approximately 150 g of seeds was dehusked and milled in a laboratory rice husker and milling machine (model JGMJ 8098, China) after cleaning the paddy with the optimal level of moisture. Three traits related to grain quality were recorded using the Standard Evaluation System in rice (http://www.knowledgebank.irri.org/images/docs/rice-standard-evaluation-system.pdf): grain aroma, head rice recovery (HRR, %), and percentage of grains with chalkiness (PGC, %). The grain aroma was estimated for each accession using a sensory method (Sood and Siddiq, 1978). Two fragrant Basmati rice varieties, *viz.*, Pusa-1121 with an aroma score of 3, PB-1 with an aroma score of 2, and a non-aromatic rice Pusa-44, were used in the analysis, and each sample was evaluated by seven experts to confirm the phenotype. Following milling, head rice recovery (HRR, %) and percentage of grains with chalkiness (PGC, %) were calculated manually and using a stereomicroscope based on the SES Scale 9, respectively. Meanwhile, the range, mean value, deviation, and phenotypic coefficient of variation (CV) were calculated for each trait using R Studio (Supplementary Table 2). Correlations of quality traits among themselves were also studied by measuring the linear correlation calculated using the R package corrr (https://cran.r-project.org/bin/windows/base/). Heritability was estimated for all three quality traits using R package variability.

Genotyping

The genomic data of 198 accessions selected from the 3K RG panel were analyzed. The SNP dataset (3K RG 1M GWAS SNP) was downloaded from the repository of rice variants in the public domain SNP-seek(http://snp-seek.irri.org/_download.zul). Missing data were imputed using Beaglev5.4 software. Quality control was performed using TASSELv5.2.82 software to obtain a filtered subset of 553,831 SNPs with a minor allele frequency >5% and a major allele frequency <95% for genome-wide association analysis.

Cluster analysis, population structure, and kinship

Neighbor-joining clustering was performed based on the SNP data using TASSELv5.2.82 software and visualized using the interactive tree of life (iTOL) software. The subgroups were assessed using a Bayesian model-based approach in STRUCTUREv2.3.4 (Pritchard et al., 2000) and PCA analysis. The structural analysis was executed with the presumed number of subgroups ranging between one and seven, with each K repeated thrice. A burin-in period of 100,000 iterations followed by 100,000 Markov Chain Monte Carlo (MCMC) simulations were implemented for every run, and the number of subgroups was then determined using the Evanno ΔK method (Evanno et al., 2005) embedded in the STRUCTURE HARVERSTER software (Earl and VonHoldt, 2012). Component analysis was performed using the Genome Association and Prediction Integrated Tool (GAPIT) R package (Lipka et al., 2012). Number of significant principal components explaining the population variance and structure were determined by plotting a scree plot in R. For kinship calculation, the Centered_identity-by-state (IBS) default method was employed in TASSELv5.2.82 software (Bradbury et al., 2007). The structure, kinship matrix, and average trait value of each accession were used for the association studies based on SNP data.

Linkage disequilibrium analysis

Linkage disequilibrium (LD) decay distance between the pair of SNP markers was calculated on each chromosome as the squared

coefficient of correlation (r^2) values of alleles using LDkit software. The position on the chromosome at which the r^2 value reduced to half of its average maximum value was defined as the decay in LD (Huang et al., 2010).

Candidate gene-based association analysis and identification of superior haplotypes

We performed GWAS on 198 rice accessions using the MLM and CMLM model with filtered 553,831 SNP markers and default settings in GAPIT software to estimate the significant SNP-MTAs for grain aroma, HRR%, and PGC%. Three multi-locus models, namely mrMLM, FASTmrMLM, and FASTmrEMMA, were also constructed using the mrMLM R package (https://cran.r-project.org/web/packages/mrMLM/index.html) to accurately detect the candidate QTN effect values and confirm the true associations. Considering an LOD score value ≥3 as the threshold, significant QTNs were identified (Duan et al., 2017). The common QTNs detected by the two different ML-GWAS models and SL-GWAS models were predicted to be good candidates for rice quality traits. Local haplotype blocks of each robust QTN were generated with all filtered SNP using PLINKv1.9 (www.cog-genomics.org/plink/1.9/) as per standard methodology (Gabriel et al., 2002). LD heatmaps were generated using the LDBlockShow tool. All genes located within the LD decay distance of the identified QTNs were extracted and subjected to comprehensive gene annotation studies to identify the candidate loci for each quality trait using The Rice Annotation Project-Database (RGAP, http://rice.uga.edu/), Information Commons for Rice (IC4R, http://ic4r.org/), and Gramene (https://www.gramene.org/) databases and used for gene mining. The haplotypes for each of these candidate loci were estimated considering the non-synonymous coding SNPs in the SNP-Seek database (https://snp-seek.irri.org/), and Student's t-test was performed to test the significant differences among the haplotypes. The haplotypes revealed and the phenotypic distribution of each grain quality trait were then represented as boxplots using the ggplot2 package in R Studio.

Results

Trait variance and correlations

Three grain quality-related traits, grain aroma, head rice recovery (HRR, %), and percentage of grains with chalkiness (PGC, %), were investigated in the selected subset of 198 accessions sampled from 3,000 re-sequenced genomes in the IRRI Rice Genome Project (3K-RGP). Rice accessions consisting of a diverse set of *Xian*, *japonica*, *aus/boro*, intermediate type cultivars, and two check varieties *viz.*, PB-1121 and PB-1 were planted at the research field of ICAR-IARI, New Delhi in 2020 and 2021. The statistical parameters were estimated, and the results are listed in Supplementary Table 2. HRR (%) and PGC (%) followed a negatively skewed distribution, whereas the grain aroma followed a positively skewed distribution (Figure 1). Furthermore, correlation analysis among the three traits indicated a statistically significant variation between the paired quality traits at the 5% and 1% levels of significance, except for the relationship between HRR (%) and PGC (%). Grain aroma was positively associated with HRR (%) (PCC = 0.28) and negatively associated with PGC (%) (PCC = −0.17), which is consistent with several previous studies (Sanchez et al., 2023; Song et al., 2007; Adjah et al., 2020; Qiu et al., 2021). In addition, HRR (%) and PGC (%) had a very weak positive correlation with a Pearson correlation coefficient (PCC) of 0.03, which was also consistent with current correlation studies and BLUP estimates (Sanchez et al., 2023; Nirmaladevi et al., 2015; Vemireddy et al., 2015; Cruz et al., 2021; Ali et al., 2023). Broad-sense heritability (H^2) estimates were high for HRR (%) (0.99) and PGC (%) (0.98) which was consistent with similar studies (Sanchez et al., 2023; Ali et al., 2023). The considerably low H^2 for grain aroma (0.28) suggested that its environmental influence was attributed to the experimental conditions, as pointed out in an earlier study (Vemireddy et al., 2015). These findings indicate a close relationship among the abovementioned quality traits and suggest their potential role in the genetic improvement of rice grain yield.

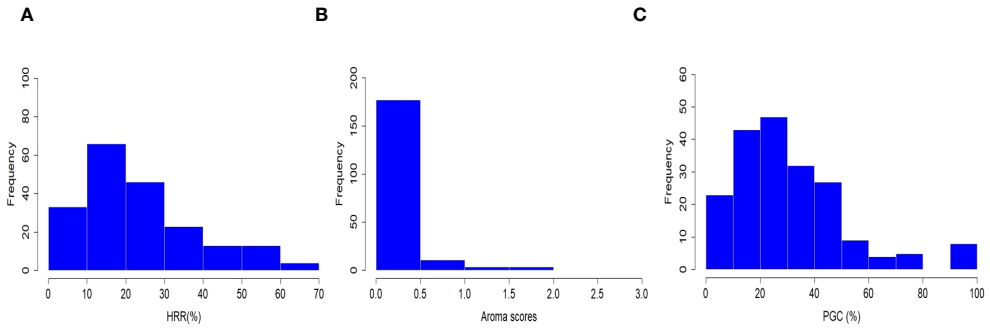


FIGURE 1
Phenotypic distribution of head rice recovery (HRR, %), grain aroma (AR), and percentage of grains with chalkiness (PGC, %) in a subset of 198 rice accessions.

Population structure and linkage disequilibrium analysis

According to principal component analysis, there were three subpopulations in the selected rice panel (Figure 2C). The scree plot suggested the significance of three PCs in the subset selected, with the first two PCs (PC1 and PC2) explaining a cumulative percent variance of 77.8 (Figure 2B). Neighbor-joining (NJ) clustering also revealed three distinct clusters based on genetic distances derived from SNP differences in the selected rice accessions (Figure 2A). Cluster 1 was identified as the smallest cluster consisting of 4.04% of *indica* rice accessions belonging to *indx* and *ind1b* subpopulations. A total of 26.26% of the *Xian* subpopulations, *viz., ind1a*, *ind1b*, *ind2*, and *ind3*, were included in cluster 2. However, cluster 3 was recognized as the largest and the most diverse cluster, comprising 69.69% of the total accessions, were *Xian*, *japonica*, *aus/boro*, and *intermediate*-type subpopulations. LD decay analysis was conducted using the filtered SNPs. Maximum r^2 estimated on the 90th percentile of chromosomes 1 to 12 was 0.3, 0.25, 0.35, 0.25, 0.35, 0.3, 0.3, 0.25, 0.3, 0.35, 0.25, and 0.25, respectively. As shown in Figure 3, variations were observed in the LD decay distance among the 12 chromosomes, with the fastest decay occurring in chromosome 12. These SNPs were found to be distributed across the whole rice genome, with an average number of SNP per kb 1.28 sufficiently dense to identify significant associations and QTLs.

Association analysis

Associations for all three traits (Aroma, HRR (%), and PGC (%)) were studied using single-locus approaches (MLM and CMLM) for QTL detection and three multi-locus methodologies (mrMLM, FASTmrMLM, and FASTmrEMMA) to identify QTNs. Using the MLM method, 198, 198, and 198 single nucleotide polymorphic (SNP) markers corresponding to 23, 22, and 32 QTLs were found to be associated with aroma, HRR (%), and PGC (%), respectively, considering the threshold value of -log (P) value = 3 (Supplementary Table 3), similar to multiple recent GWAS studies (Kikuchi et al., 2017; Bheeanahalli et al., 2021; Hu et al., 2022). Of these, 24 QTNs for aroma using mrMLM (10), FASTmrMLM (11), and FASTmrEMMA (3). For HRR (%), eight, 11, and four QTNs were detected using mrMLM, FASTmrMLM, and FASTmrEMMA, respectively, and 23 QTNs were correlated with PGC (%) using mrMLM and FASTmrMLM (Supplementary Figure 1). Manhattan and quantile-quantile plots of all the three quality traits presented in Figure 4 implied that false associations were controlled and the SNPs detected by ML-WAS methods were true associations; however, we witnessed inflation in Q–Q plots with incorporated population structure. This inflation persisted because the mixed linear approach (accounting for structure) utilized the first three PCs as covariates in the regression. However, the PC-adjusted model-based estimates of standard errors remove the structure problem, providing correctly calibrated p-values, which has been well documented in several studies (Price et al., 2006; Zhang et al., 2008; Voorman et al., 2011). One of the QTNs detected for aroma (qAR-1-1) was located in proximity to the well-known rice fragrance gene *Badh2* (151 kb). The recessive gene *BADH2* is well established to govern the synthesis of 2-acetyl-1-pyroline (2-AP) in aromatic rice (Imran et al., 2022). Furthermore, we found that *qHRR-3-1* existed in the same region adjacent to *OsRLCK113* (cysteine-rich receptor-like kinase 28 precursor gene, *LOC_Os03g31260*) (Li et al., 2022) and the gene encoding the ring zinc finger protein (*LOC_Os03g31320*) (65–

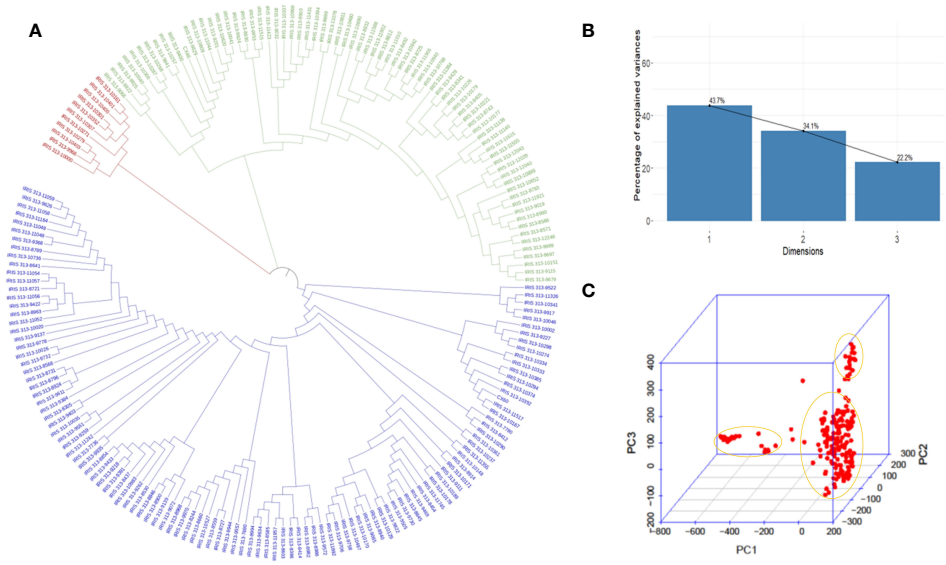


FIGURE 2
Model and PCA based analysis of genetic structure of 198 rice accessions. **(A)** NJ clustering of 198 rice accessions constructed using 5,53,229 SNPs evenly distributed throughout the genome. **(B)** Scree plot depicting the genetic variation with principal components. PC1, PC2, and PC3 represent the first, second, and third principal components, respectively. **(C)** Biplot depicting three clusters identified in the selected rice panel.

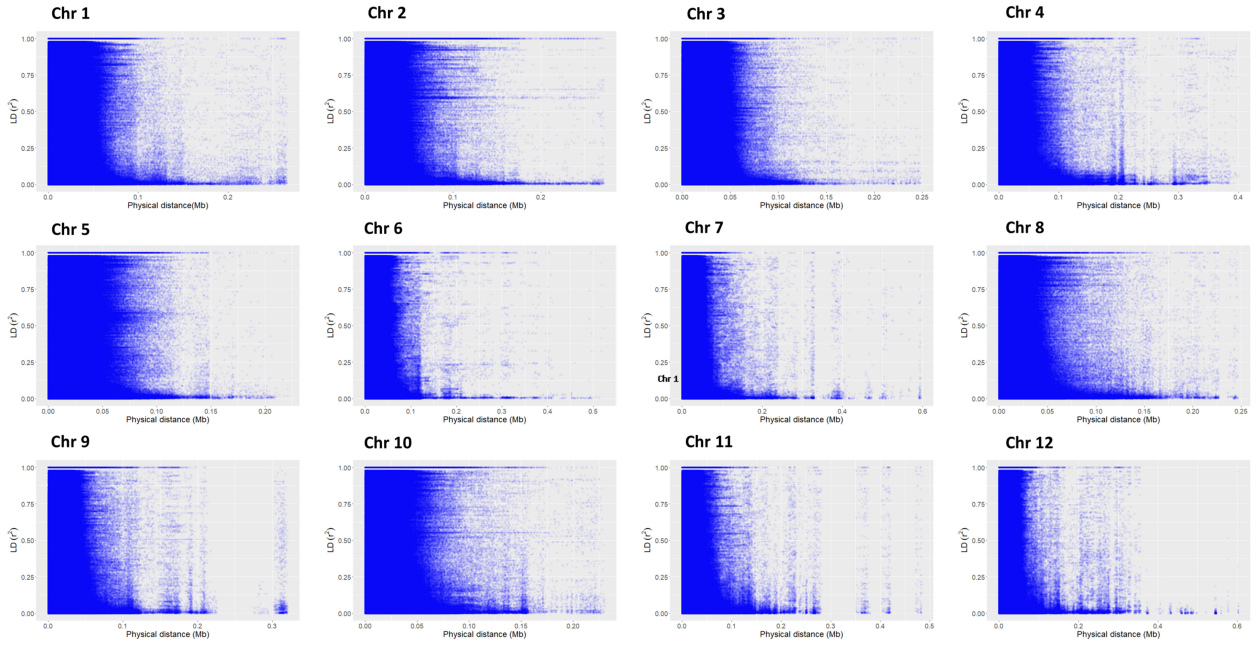


FIGURE 3
Chromosome-wise linkage disequilibrium decay based on 198 accessions. The decline in LD-r^2 between SNP markers is presented as a function of physical distance in base pairs.

66), with confirmed roles in controlling grain yield and quality traits in rice. *qPGC-3-1* and *qPGC-3-2* were located adjacent to *OsLTP1.3* (Ltpl28-Seed Storage/Protease Inhibitor/Ltp Family Protein Precursor, *LOC_03g59380*), *OsCESA2* (Cellulose Synthase, *LOC_03g59340*) and *OsCPK8* (Camk_Camk_Like.24- Calcium Dependent Protein Kinases, *LOC_03g59390*) genes regulating grain quality traits in rice. Similarly, *qPGC-7-2* overlapped with a gene encoding a retrotransposon protein located in the vicinity of no apical meristem genes *ONAC65* (*LOC_07g27330*) and *ONAC102* (*LOC_07g27340*), which serves as a regulator of starch and accumulation of proteins, thereby improving grain quality in rice (Wang et al., 2020).

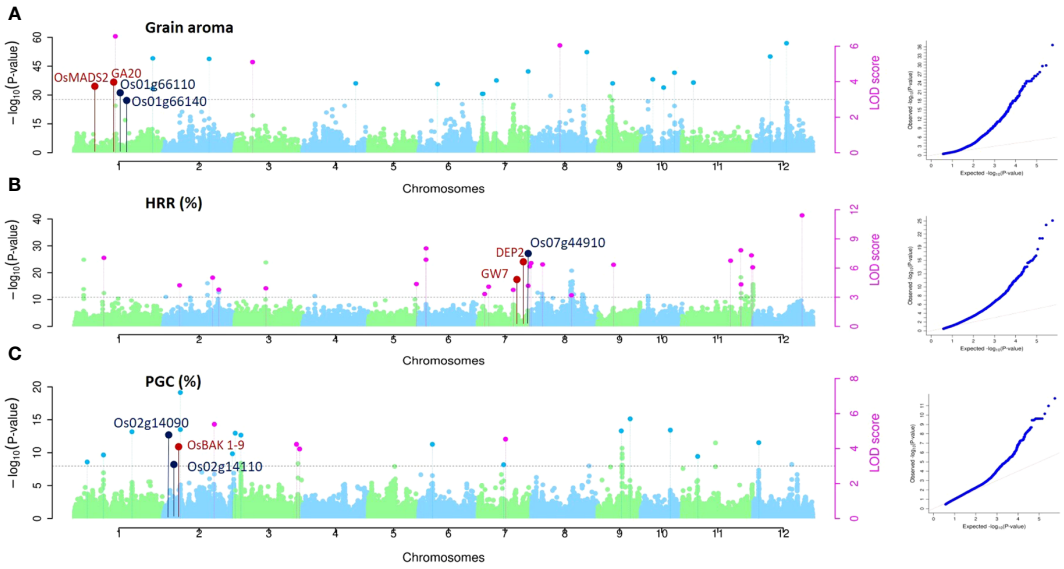


FIGURE 4
GWAS for grain quality traits in rice accessions. Manhattan and Quantile-Quantile plots derived through the mrMLM, FASTmrMLM, and FASTmeEMMA methods depicting the distribution of QTNs on 12 rice chromosomes for grain aroma (AR), head rice recovery HRR (%), and percentage with grain chalkiness PGC (%). Pink dots indicate all the QTNs mapped by more than one GWAS method, while all the QTNs identified by a single method are indicated by the light colored dots shown above the gray dotted lines. The known genes around QTNs are marked in red, and putative candidates around the identified QTNs are marked in dark blue.

10.3389/fpls.2023.1304388

In addition, assessment of the results of SL-GWAS and ML-GWAS revealed 39 QTNs in common based on a critical LOD score ≥3, explaining 0.03%–9.57% of the phenotypic variation (R^2) (Table 1). The fact that half of the detected QTNs (19/39) overlapped with previously reported genes/QTLs supports the consistency of our results. Among these, 7, 13, and 19 were associated with AR, HRR (%), and PGC (%), respectively. Seven candidate QTNs significantly related to AR were located on chromosomes 1, 3, 8, 10, 11, and 12. For HRR (%), 13 putative QTNs were found to be distributed on chromosomes 2, 3, 6, 7, 8, and 11. A total of 19 QTNs correlated with PGC (%) were found to be located on chromosomes 1, 2, 3, 5, 7, 9, 10, 11, and 12. Of these,

TABLE 1 QTNs for the three quality traits detected concurrently by using single- and multi-locus GWAS methodologies.

Trait	QTN	Chr	Position	LOD	R^2(%)[1]	Method[2]	LOC[3]/QTL[4]
Aroma (AR)	qAR-1-1	1	20227999	0.05–9.03	0.05–7.5	1,2,3,4,5	
	qAR-1-2	1	38383904	3.59	3.16	1,2,3	GA20ox-2
	qAR-3-1	3	10338993	4.46–5.74	0.13–5.89	1,3,4	
	qAR-8-1	8	10892476	5.54–6.54	4.25–4.93	3,4	LTP48/CQAP1
	qAR-10-1	10	17265187	4.503	0.09–2.57	1,2,5	OsCESA7
	qAR-11-1	11	6394202	3.9514	0.03–6.53	1,2,3	OsSRP-PLP
	qAR-12-1	12	15819670	6.1642	0.09–3.94	1,2,5	CQAP3
Head Rice Recovery HRR (%)	qHRR-2-1	2	24752396	5.0096	0.05–5.71	1,2,5	hwh1, AQCV031[a]
	qHRR-3-1	3	17840988	3.9148	0.09–4.47	1,2,4	LOC_Os03g31310
	qHRR-6-1	6	3667482	6.8549	0.08–8.20	1,2,3	
	qHRR-6-2	6	3730045	8.0183	0.09–5.51	1,2,3	
	qHRR-7-1	7	20413747	3.7534	0.05–3.97	1,2,5	LOC_Os07g34130
	qHRR-7-2	7	26771672	4.1753	0.06–4.47	1,2,5	LOC_Os07g44830
	qHRR-7-3	7	28019959	6.168	0.06–6.95	1,2,3	
	qHRR-8-1	8	4580996	6.3682	0.05–7.59	1,2,5	
	qHRR-8-2	8	16979079	3.2005	0.07–2.00	1,2,4	
	qHRR-11-1	11	21623134	6.7537	0.06–4.22	1,2,4	LOC_Os11g36640
	qHRR-11-2	11	24456311	7.8223	0.06–7.02	1,2,4	
	qHRR-11-3	11	27996997	7.3025	0.07–4.78	1,2,4	OsPCBP
	qHRR-11-4	11	28857401	6.0719	0.05–5.69	1,2,3	OsRhmbd18
Percentage with grain chalkiness PGC (%)	qPGC-1-1	1	14474816	3.6322	3.59	1,2,3	LOC_Os01g25530
	qPGC-1-2	1	5928150	3.2289	5.51	1,2,3	LOC_Os01g11110
	qPGC-2-1	2	25081182	3.54–7.22	8.45–9.57	1,2,3,4	LOC_Os02g41720
	qPGC-2-2	2	7633393	7.19	6.15	1,2,3	LOC_Os02g13990, AQGB108[b]
	qPGC-2-3	2	7660595	5.08	2.95	1,2,4	AQGB109[c], AQGB084[d]
	qPGC-2-4	2	34652183	3.69	1.55	1,2,4	LOC_Os02g56565
	qPGC-3-1	3	35098972	3.75–4.18	2.08–4.98	1,2,3,4	
	qPGC-3-2	3	33802678	3.99–4.48	2.8-6.1	1,2,3,4	
	qPGC-3-3	3	326950	4.87	3.52	1,2,3	
	qPGC-3-4	3	4530119	4.76	3.34	1,2,4	
	qPGC-5-1	5	6812458	4.23	8.51	1,2,3	
	qPGC-7-1	7	620874	6.4849	0.07–1.14	1,2,3	
	qPGC-7-2	7	15975911	3.07	2.35	1,2,4	LOC_Os07g27420

(Continued)

TABLE 1 Continued

Trait	QTN	Chr	Position	LOD	R^2(%)[1]	Method[2]	LOC[3]/QTL[4]
	qPGC-7-3	7	16539429	3.9–5.16	3.9–5.37	1,2,3,4	
	qPGC-9-1	9	11755843	5.0082	2.7252	1,2,4	
	qPGC-9-2	9	16443727	5.696	3.528	1,2,4	
	qPGC-10-1	10	14524700	5.05	3.21	1,2,4	
	qPGC-11-1	11	8517398	3.5495	2.8693	1,2,4	
	qPGC-12-1	12	3125286	4.3352	3.4133	1,2,4	

[1]R^2(%): phenotypic variance explained.
[2]Methods 1–5 represent MLM, CMLM, mrMLM, FASTmrMLM, and FASTmrEMMA, respectively.
[3]Locus name based on MSU 7.0.
[4]QTL ID based on Gramene QTL Database. [a]Li et al. (2003); [b,c,d]Wan et al. (2005).

four QTNs were detected by both SL-GWAS and at least two ML-GWAS methods (qPGC-2-1, qPGC-3-1, qPGC-3-2, and qPGC-7-3). As many as 75 cloned genes were closely associated with rice yield and appearance quality within the genomic ranges (± 100 kb) of the 39 QTNs detected by the SL-GWAS and ML-GWAS methods (Figure 5; Supplementary Table 4).

Mining of potential candidate loci

We selected common QTNs mapped using the SL-GWAS and ML-GWAS algorithms for a detailed study. The candidate genes were identified based on haplotype analysis of non-synonymous coding SNPs in each candidate gene located inside the LD block defined for the selected QTN.

qAR-1-2, located at 38,383,904bp on chromosome 1, showed association signals with grain aroma using MLM, CMLM, and mrMLM methods with a Logarithm of Odds (LOD) score of 3.59% (Table 1). A total of 54 kb LD block (38,375,000 bp–38,429,000 bp) was generated (Figure 6A) as per the method described above (Gabriel et al., 2002). Gene annotations suggested five candidates for this block: *LOC_Os01g66100* (gibberellin20oxidase2 gene, *OsGA20ox2*), *LOC_Os01g66110* (a methyltransferase), *LOC_Os01g66120* (no apical meristem protein-encoding gene, *OsNAC6*), *LOC_Os01g66130* (an armadillo/beta-catenin repeat family protein, *OsPUB16*), and *LOC_Os01g66140* (plus-3 domain-containing protein). Among these, *LOC_Os01g66110* is the most likely gene because the heavy metal transporter genes involved in the biosynthesis of 2-AP, which determines the aroma in fragrant rice, are known to be regulated by DNA methylases via active histone modifications (Imran et al., 2022). Missense mutations in *LOC_Os01g66110* resulted in three allelic combinations. Genotypes with superior HapA exhibited higher average aroma scores, whereas genotypes with HapB and HapC showed lower aroma scores (Figure 6B). Another candidate gene, *LOC_Os01g66140*, directly interacts with histone H4 and zinc ions, explaining its role in 2-AP biosynthesis. Three haplotypes were observed for *LOC_Os01g66140*, and haplotype A showed a significantly higher average aroma score than the other two haplotypes.

The SL-GWAS and ML-GWAS test results verified peaks on chromosome 7 for HRR (%). qHRR-7-2, located at 26,771,672 bp and encoding a proline-rich family protein, was significantly linked to HRR (%) with the FASTmrEMMA method with an LOD score of 4.17 (Table 1). Using MLM, this SNP showed associations with HRR (%) with a high level of significance ($p = 3.84 \times 10^{-6}$) and an R^2 of 5.43%. An LD block of 26,760,000 bp to 26,798,000 bp was constructed using pairwise estimation of LD (Figure 7A). The fine mapping of this genetic region associated with HRR (%) identified five candidate genes using genome annotation tools: *LOC_Os07g44830* belonging to the proline-rich family, *LOC_Os07g44840* encoding a transposon with unknown function, and *LOC_Os07g44850*, *LOC_Os07g44860*, *LOC_Os07g44900*, and *LOC_Os07g44910* are gibberellin receptor protein-encoding genes. The *LOC_Os07g44910*, annotated as putatively expressed gibberellin receptor GID1L2 protein, showed significant differences in HRR (%) between the haplotypes (Figure 7B). Therefore, HapA is a superior genotype and rice accessions with a higher frequency of HapA could be selected from the current panel to improve head rice recovery (%) in rice. Earlier studies clearly indicated the role of Gibberellic Acid in controlling panicle architecture and yield traits in rice (Deveshwar et al., 2020). Moreover, *LOC_Os07g44910* colocalized with the *dense and erect panicle 2* (*DEP2*) gene, which is mainly involved in rachis elongation and branching in panicles (Li et al., 2003; Wan et al., 2005), and the *GW7* gene, which encodes a TONNEAU1-recruiting motif protein that improves grain yield and quality by directly interacting with *GW8* (*OsSPL16*) (Li et al., 2010; Reig-Valiente et al., 2018). We utilized the IC4R database to confirm the functional role and analyzed the expression profile data of *LOC_Os07g44910* in rice and found that the gene encodes an alpha/beta hydrolase fold-3 domain-containing protein with the highest expression in the seedlings and young shoots. Previous studies have reported that the *D14* gene encoding an alpha/beta hydrolase family protein inhibits rice tillering via the strigolactone signaling pathway (Gao et al., 2009; Wang et al., 2015; Guo et al., 2020); thus, it is likely that *LOC_Os07g44910* influences grain yield in rice.

qPGC-2-3 was another QTN detected by multiple models and showed associations with the percentage of grains using chalkiness FASTmrMLM methods with an LOD value of 5.08. This QTN was also detected by the MLM and CMLM methods with a p value of 3.05×10^{-6}. An LD block was defined for this QTN (83.63 kb), and

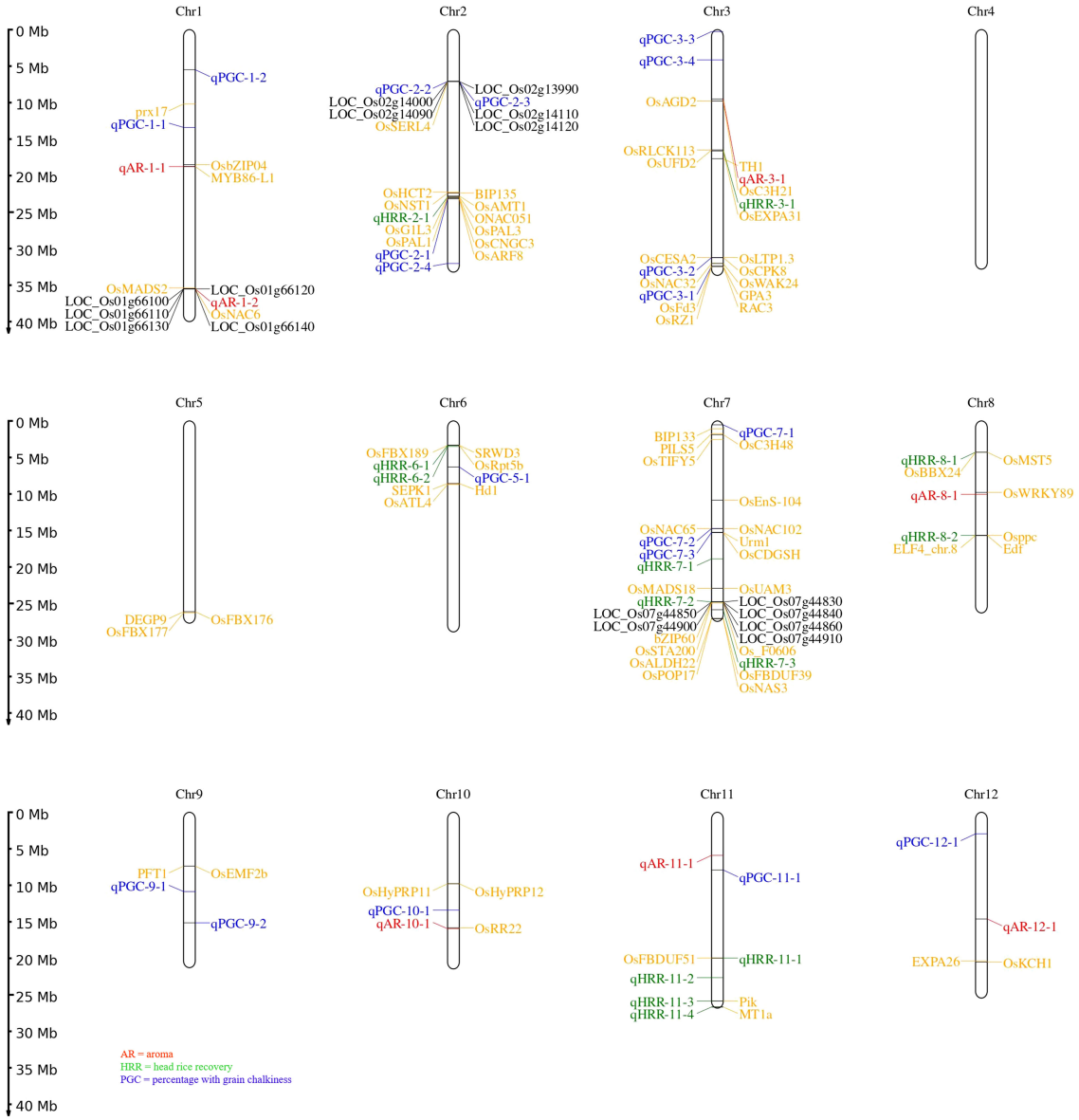


FIGURE 5
Chromosomal distribution of all loci for grain quality traits using MLM, CMLM, mrMLM, FASTmrMLM, and FASTmeEMMA. The naming of QTNs starts with a letter 'q' subsequently followed by two or three letter identifiers and the chromosome number. In case numerous QTNs are mapped for a quality trait on corresponding chromosome at that point naming is done based on their relative location on the chromosome. Seventy-five known genes are labelled with yellow script; black color represents candidate genes for the quality traits under study.

five candidate genes were identified in this region (Figure 8A). *LOC_Os02g13990* (U2 small nuclear ribonucleoprotein A) and *LOC_Os02g14000* (actin-related protein 2/3 complex subunit 3) only had synonymous SNPs with a -log (P) value less than 3. *LOC_Os02g14120* is a Brassinosteriod Insensitive 1 Associated Receptor Kinase 1 precursor gene (*OsBAK 1-9*). Non-synonymous mutations in *OsBAK 1-9* resulted in three major haplotypes: HapA, HapB, and HapC. The accessions with favorable HapA displayed lower PGC (%) than accessions with HapB and HapC types (Figure 8B). The identified favorable allele and functional site in *LOC_Os02g14120* reduces the degree of chalkiness in rice by breeding. Differences in rice grain quality have been attributed to the regulation by a set of other genes involved in multiple pathways that influence grain appearance quality. *LOC_Os02g14110* is annotated as an aminotransferase, Class I and Class II domain-containing protein gene, and the third candidate gene, *LOC_Os02g14090*, is a berberine and berberine-like domain-containing protein gene. Previous research has also verified that brassinosteroid-associated receptor kinase genes, putatively expressed aminotransferases, and berberine and berberine domain-containing protein genes govern quality traits, *viz.*, chalkiness and grain shape (Biselli et al., 2015) in rice, which led us to hypothesize that *LOC_Os02g14120*, *LOC_Os02g14110*, and *LOC_Os02g14090* may be rice grain PGC (%) regulatory genes.

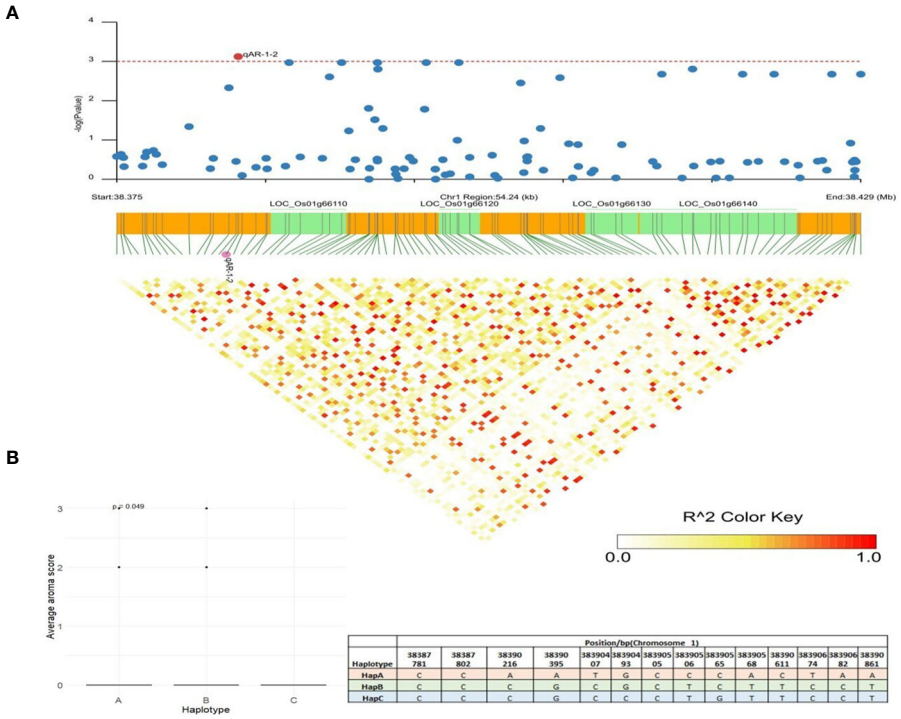


FIGURE 6
Haplotype analysis of LOC_Os01g66110. **(A)** Linkage disequilibrium (LD) based heatmap for qAR-1-2. **(B)** Boxplot for grain aroma depicting three allelic combinations of LOC_Os01g66110. X-axis shows the three different alleles of LOC_Os01g66110 and Y-axis shows the average aroma scores.

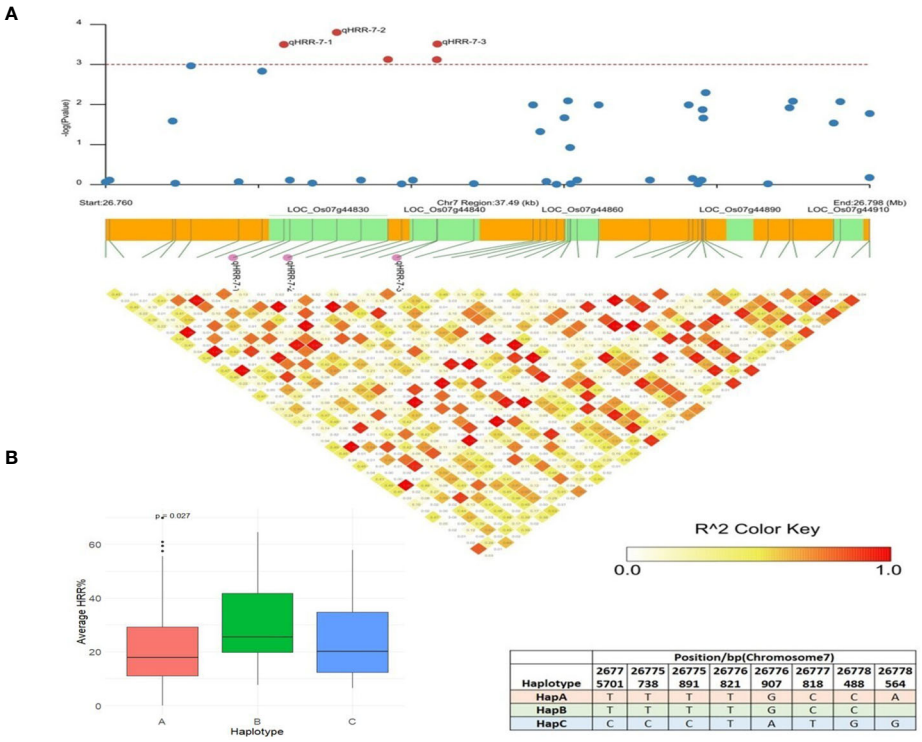


FIGURE 7
Haplotype analysis of LOC_Os07g44910. **(A)** Linkage disequilibrium (LD) based heatmap for qHRR-7-2. **(B)** Boxplot of HRR (%) trait depicting three allelic combinations of LOC_Os07g44910. The X-axis shows three different alleles of LOC_Os07g44910 and the Y-axis shows average HRR (%).

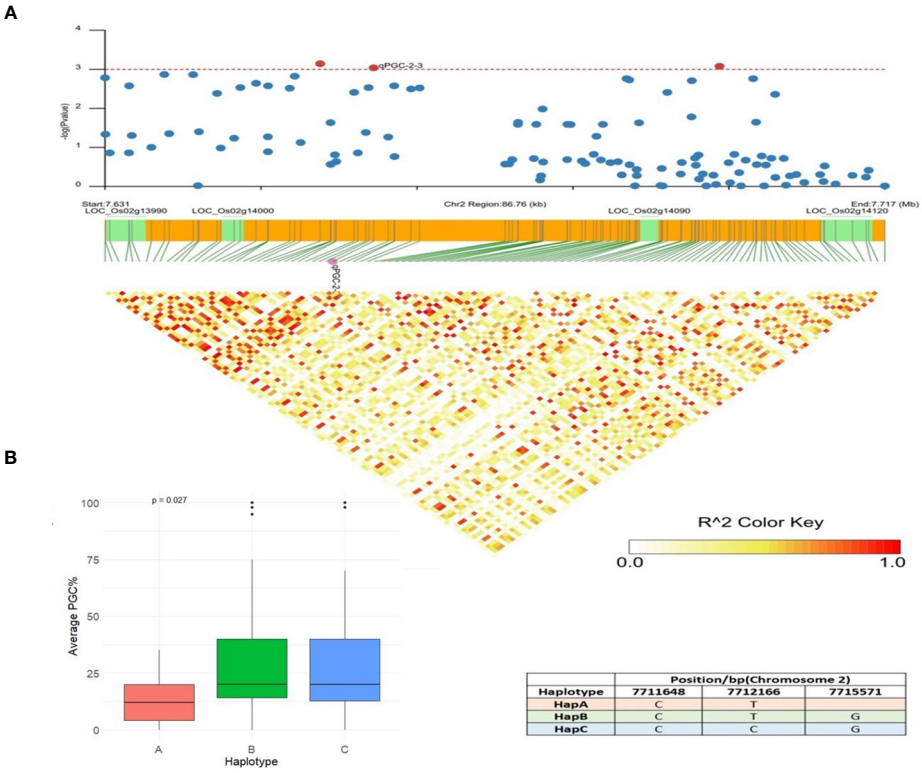


FIGURE 8
Haplotype analysis of LOC_Os02g14120. **(A)** Linkage disequilibrium (LD) based heatmap for qPGC-2-3. **(B)** Boxplot of PGC (%) trait depicting three allelic combinations of LOC_Os02g14120. X-axis shows three different alleles of LOC_Os02g14120 and Y-axis shows average PGC (%).

Discussion

Increasing living standards underline the need to develop healthier high-quality rice (Yu et al., 2013; Hori et al., 2015; Sahu et al., 2017; Wang et al., 2017; Misra et al., 2019; Meng et al., 2022) for traits such as color, aroma, lack of broken seed grains, grain length, and flavor. To meet consumer preferences and market demands, the development of tailored rice with preferred appearance quality is of utmost importance after rice yield enhancement (Arite et al., 2009; Abbai et al., 2019; Selvaraj et al., 2021). Grain quality is a complex quantitative trait (Yu et al., 2013; Hori et al., 2015; Misra et al., 2019; Meng et al., 2022) governed by manifold genes, and there is a large gap in our perception of the networks regulating grain quality in rice (Li et al., 2022). GWAS has become a robust tool for the rapid identification of genetic factors (Adjah et al., 2020) associated with traits governed by several genes in crop plants that are diverse and provides goals for future efforts aimed at rice improvement (Zhou et al., 2020). However, breeding by design has achieved limited success because of the lack of information on the correct genetic loci of desired traits and precision in deciphering the favorable haplotype combinations of these genes dissected to date (Fitzgerald et al., 2009; Abbai et al., 2019; Selvaraj et al., 2021).

Resequencing-based germplasm lines enable the detection of pre-existing variations, functional sites of genes, and novel alleles associated with traits of interest (Begum et al., 2015), which may be explored by GWAS analysis. In this context, the abundant genetic variations in 3K RG resequencing projects make it a valuable reservoir of gene diversity and a prospective source of elite genes that can be deployed in rice breeding (Abbai et al., 2019; Selvaraj et al., 2021). Traditional single-locus models, which are commonly adopted to identify genetic variants in several cereal crops, have some limitations, neglecting small-effect QTLs in particular. Lower false positives and higher statistical predictions of multi-locus algorithms have been established by many association studies (Yuan et al., 2017; Zhang P. et al., 2019), and researchers usually combine facts about different ML-GWAS models to mine the genes that control complex traits.

In the present study, we adopted two SL-GWAS methods and three ML-GWAS methods to assess three quality traits of 198 selected rice accessions (a subset of 3K RGP). Subsequently, 198,198 and 198 significant SNPs, while 23,22 and 32 QTLs were identified by MLM underlying AR, HRR (%), and PGC (%), respectively (Supplementary Table 3). Similarly, 24,23 and 23 significant QTNs were detected using ML-GWAS methodologies associated with the abovementioned three quality traits (Supplementary Figure 1). Interestingly, the QTNs mapped by multi-locus GWAS analysis were more dispersed than those mapped by the MLM and CMLM methods. The significant loci detected by the MLM method, for example, were confined to specific chromosomes, indicating its failure to identify new loci across the entire rice genome. Several QTNs identified by multi-locus methods were distributed across the other chromosomes, among which 39 common QTNs were considered powerful, robust,

and worthy when applied to discover low individual QTN effect values for quality traits in rice.

Several rice grain quality genes, such as *Badh2*, *DEP2*, *GW7*, *OsCESA2*, and *OsCPK8*, have been functionally characterized over the past 10 years (Deveshwar et al., 2020; Imran et al., 2022; Yan et al., 2022). Among these, *Badh2*, the *fgr* gene, the major gene causing fragrance in rice and a frameshift mutation in its exonic region, is the functional allele associated with fragrance (Quero et al., 2018; Tibbs et al., 2021). *DEP2/SRS1* encoding the dense and erect panicle 2 gene positively regulates panicle morphology and its outgrowth, suggesting its direct role in regulating rice grain size and yield at the genetic level (Li et al., 2010). *GW7* is annotated as a gene encoding a TONNEAU1-recruiting motif protein that simultaneously controls grain width and quality (Li et al., 2010).

Combining the cloned genes/QTLs reported in earlier genetic studies, 19 QTNs and their ±100 kb genomic regions superimposed the previously annotated grain-quality genes. QTNs clusters were mapped for HRR (%) on chromosome 7 (*qHRR-7-1*, *qHRR-7-2*, and *qHRR-7-3*) located in the vicinity of *GW7* and *DEP2*, which are responsible for grain yield and quality, and another cluster was detected on chromosome 11 (*qHRR-11-1*, *qHRR-11-2*, *qHRR-11-3*, and *qHRR-11-4*) near the F-box and DUF domain-containing genes with confirmed roles in improving yield potential and quality in rice. Additionally, 20 novel QTNs were excluded from the genomic loci of earlier studies, and the markers detected may be the putative QTNs governing quality traits in rice.

Dissecting four candidate genes of grain quality traits

Using multiple models for association studies, three QTNs (*qAR-1-2*, *qHRR-7-2*, and *qPGC-2-3*) were confirmed to have major gene effects on grain quality. The candidate region of 38.37 Mb to 38.42 Mb in qAR-1-2 was fine-mapped considering a threshold value of r^2 >0.2 (Figure 6A). Five genes located in this genomic region were possible candidates governing aroma in rice, and haplotyping was performed for each of the five genes. Significant differences in aroma scores between the *LOC_Os01g66110* and *LOC_Os01g66140* haplotypes were observed (Figure 6B). *LOC_Os01g66110*, a putative methyltransferase, has been proposed to play a role in multiple epigenomic modifications of heavy-metals transporters involved in the 2-AP biosynthesis pathway. In recent years, the occurrence of DNA methylation of all types (CHH, CHG, and CG) in genes related to 2-AP biosynthesis has been reported in rice. ChIP-seq, bisulfite-seq, and ATAC-seq data of aroma genes also showed active chromatin modifications as key regulators (Imran et al., 2022) with strong enrichment of H3K36me3 at 2-AP biosynthesis pathway-related genes. Another candidate gene, *LOC_Os01g66140*, annotated as a plus-3 domain-containing protein, is anticipated to influence 2-AP biosynthesis genes with metal-binding properties and DNA-binding domains. BLAST tool and STRING analysis revealed that *LOC_Os01g66140* directly interacts with histone H4 and zinc metal ions, confirming its role in regulating 2-AP content in aromatic rice. Prior studies have found that exogenous application of micronutrients, specifically zinc, could upregulate genes involved in the biosynthesis of 2-AP in aromatic rice due to increased levels of proline and proline dehydrogenase (He and Park, 2015). Based on these findings, we propose that *LOC_Os01g66110* and *LOC_Os01g66140* may be related to the grain aroma. Their role in regulating heavy metal transporters in response to zinc is worthy of comprehensive studies and confirmation.

The candidate *qHRR-7-2* associated with HRR (%), *LOC_Os07g44910*, annotated as the gibberellin receptor *GID1L2*, is a type of F-box subunit of the S-phase kinase-associated protein 1 (SKP1)-cullin 1 (CUL1)-F-box protein (SCF) E3 complex that encodes the alpha/beta hydrolase fold-3 domain-containing protein containing 358 amino acids, belonging to the alpha/beta hydrolase (ABH) superfamily. The F-box protein (SCF) E3 complex plays a crucial role in regulating life processes such as cell division and influences grain size and yield in rice by facilitating proteasomal degradation of diverse regulatory proteins (Chen et al., 2008; Nguyen and Businо, 2020). Its loss-of-function mutants, htd4 and dta-34 have reduced panicle branching, grains/panicle, and seed size, and show a dwarf phenotype (Wang et al., 2017; Liang et al., 2019, Liu et al, 2009). For instance, *Grain weight 2* (*GW2*), encoding E3 ubiquitin ligase, regulates grain weight and grain yield by ubiquitinating *EXPLA 1* and promoting its degradation (Hu et al., 2015; Mo et al., 2016; Deveshwar et al., 2020). In this study, GWAS and haplotype analysis results indicated that *LOC_Os07g44910* might govern grain weight and grain yield in rice (Figures 7A, B). Members of this superfamily, such as *GS5* (*Grain Size 5*, putative serine carboxypeptidase) (Hu et al., 2015) and *TGW6* (*Thousand Grain Weight 6*, IAA-glucose hydrolase) (Mo et al., 2016), have been characterized for their roles in influencing grain weight and yield. These studies showed high consistency with our GWAS analysis results, confirming with these printed reports proving that *LOC_Os07g44910* might be related to rice recovery % (HRR, %) in rice.

In the candidate *qPGC-2-3*, involved in the percentage of grains with chalkiness, *LOC_Os02g14120* is a Brassinosteriod Insensitive 1 Associated Receptor Kinase 1 precursor gene (*OsBAK 1-9*). *OsBAK1*/*Top Bending Panicle 1* encodes a somatic embryogenesis receptor kinase (SERK) domain-containing protein that acts as a modulating factor in the brassinosteroid signaling pathway, thus affecting the number of grains and yield in rice (Xing and Zhang, 2010; Gupta et al., 2022). Overexpression of OsBAK-1 drastically reduced grain yield in rice (Lin et al., 2017), and its high-tillering mutants are characterized by a reduction in panicle length and seed size (Deveshwar et al., 2020). The central role of brassinosteroids (BR) in regulating multiple biological processes such as flowering, male fertility, and tillering, is becoming more apparent (Lin et al., 2017; Yuan et al., 2017). Although, brassinosteroids have been demonstrated to be positive regulators of plant growth processes and grain development, they most often work in close association with auxins and cytokinins to affect the efficiency of photosynthesis, sugar metabolism, and mobilizing resources in crop plants to influence grain filling (Mo et al., 2016; Deveshwar et al., 2020), reiterating the need to consider the holistic approach of plant

developmental processes and their architecture to improve crop yields. These results suggest that *LOC_Os02g14120* may be related to the percentage of grains with chalkiness (PGC, %), and its role in modulating the architecture, yield, and grain quality in rice is valuable for further evaluation and validation.

Conclusions

In this GWAS analysis, 70 QTNs were detected for three grain quality traits using different multi-locus methodologies. Among these QTNs, *qAR-1-2*, *qHRR-7-2*, and *qPGC-2-3*, which are closely associated with AR, HRR (%), and PGC (%), were identified using both single- and multi-locus methods. In addition, four key annotated genes (*LOC_Os01g66110*, *LOC_Os01g66140*, *LOC_Os07g44910*, and *LOC_Os02g14120*) that govern the three target candidate genes mentioned above were mined. In conclusion, several robust QTLs and four candidate functional genes were shown to possibly control grain aroma, head rice recovery (%), and the percentage of grains with chalkiness in rice. This investigation provides valuable information for functional characterization in the future and molecular marker-based breeding design to improve appearance quality traits in rice.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Files, further inquiries can be directed to the corresponding author/s.

Author contributions

SS: Data curation, Formal Analysis, Methodology, Software, Validation, Writing – original draft. RS: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Supervision, Visualization, Writing – review & editing. AM: Formal Analysis, Software, Writing – review & editing. VS: Writing – review & editing. US: Resources, Writing – review & editing. AK: Writing – review & editing. GS: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by Department of Biotechnology Government of India, grant number "BT/PR32853/AGIII/103/1059/2019". The funder has no role in the design of the study and data collection, analysis, interpretation and in writing the manuscript.

Acknowledgments

We would like to thank the International Rice Research Institute South Asia Regional Centre (IRRI-SARC), Varanasi, India for providing the rice accessions for this study, and Dr. A. K. Singh, Director, ICAR-Indian Agricultural Research Institute for providing the agricultural land for conducting the experiments.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1304388/full#supplementary-material

References

Abbai, R., Singh, V. K., Nachimuthum, V. V., Sinha, P., Selvaraj, R., Vipparla, A. K., et al. (2019). Haplotype analysis of key genes governing grain yield and quality traits across 3K RG panel reveals scope for the development of tailor-made rice with enhanced genetic gains. *Plant Biotechnol. J.* 17 (8), 1612–1622. doi: 10.1111/pbi.13087

Adjah, K. L., Abe, A., Adetimirin, V. O., and Asante, M. D. (2020). Genetic variability, heritability and correlations for milling and grain appearance qualities in some accessions of rice (*Oryza sativa* L.). *Physiol. Mol. Biol. Plants* 26, 1309–1317. doi: 10.1007/s12298-020-00826-x

Ali, F., Jighly, A., Joukhadar, R., Niazi, N. K., and Al-Misned, F. (2023). Current status and future prospects of head rice yield. *Agriculture* 13 (3), p.705. doi: 10.3390/agriculture13030705

Alqudah, A. M., Sallam, A., Baenziger, P. S., and Börner, A. (2020). Gwas: Fast-Forwarding gene identification and characterization in temperate cereals: Lessons from barley–A review. *J. Adv. Res.* 22, 119–135. doi: 10.1016/j.jare.2019.10.013

Arite, T., Umehara, M., Ishikawa, S., Hanada, A., Maekawa, M., Yamaguchi, S., et al. (2009). d14, a strigolactone-insensitive mutant of rice, shows an accelerated outgrowth of tillers. *Plant Cell Physiol.* 50 (8), 1416–1424. doi: 10.1093/pcp/pcp091

Bao, J. (2014). Genes and QTLs for rice grain quality improvement. In W. Yan and J. Bao Eds. *InTech–Open Science. Open Mind*, pp. 239–278. doi: 10.5772/56621

Bao, J. (2019). "Rice milling quality," in *Rice* (*4th ed.*) (St, Paul, MN: AACC International Press), 339–369. doi: 10.1016/B978-0-12-811508-4.00010-1

Begum, H., Spindel, J. E., Lalusin, A., Borromeo, T., Gregorio, G., Hernandez, J., et al. (2015). Genome-wide association mapping for yield and other agronomic traits in an elite breeding population of tropical rice (*Oryza sativa*). *PloS One* 10 (3), e0119873. doi: 10.1371/journal.pone.0119873

Bevan, M. W., Uauy, C., Wulff, B. B., Zhou, J., Krasileva, K., and Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature* 543 (7645), 346–354. doi: 10.1038/nature22011

Bheeanahalli, R., Knight, M., Quinones, C., Doherty, C. J., and Jagadish, S. K. (2021). Genome-wide association study and gene network analyses reveal potential candidate genes for high night temperature tolerance in rice. *Sci. Rep.* 11, 6747. doi: 10.1038/s41598-021-85921-z

Biselli, C., Bagnaresi, P., Cavalluzzo, D., Urso, S., Desiderio, F., Orasen, G., et al. (2015). Deep sequencing transcriptional fingerprinting of rice kernels for dissecting grain quality traits. *BMC Genomics* 16 (1), 1–28. doi: 10.1186/s12864-015-2321-7

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Champagne, E. T. (2008). Rice aroma and flavor: A literature review. *Cereal Chem.* 85 (4), 445–454. doi: 10.1094/CCHEM-85-4-0445

Chan-In, P., Jamjod, S., Yimyam, N., Rerkasem, B., and Pusadee, T. (2020). Grain quality and allelic variation of the Badh2 gene in Thai fragrant rice landraces. *Agronomy* 10 (6), 779. doi: 10.3390/agronomy10060779

Chen, S., Yang, Y., Shi, W., Ji, Q., He, F., Zhang, Z., et al. (2008). *Badh2*, encoding betaine aldehyde dehydrogenase, inhibits the biosynthesis of 2-acetyl-1-pyrroline, a major component in rice fragrance. *Plant Cell* 20 (7), 1850–1861. doi: 10.1105/tpc.108.058917

Contreras-Soto, R. I., Mora, F., de Oliveira, M. A. R., Higashi, W., Scapim, C. A., and Schuster, I. (2017). A genome-wide association study for agronomic traits in soybean using SNP markers and SNP-based haplotype analysis. *PloS One* 12 (2), e0171105. doi: 10.1371/journal.pone.0171105

Cruz, M., Arbelaez, J. D., Loaiza, K., Cuasquer, J., Rosas, J., and Graterol, E. (2021). Genetic and phenotypic characterization of rice grain quality traits to define research strategies for improving rice milling, appearance, and cooking qualities in Latin America and the Caribbean. *Plant Genome* 14 (3), e20134. doi: 10.1002/tpg2.20134

Deveshwar, P., Prusty, A., Sharma, S., and Tyagi, A. K. (2020). Phytohormone-mediated molecular mechanisms involving multiple genes and QTL govern grain number in rice. *Front. Genet.* 11, 586462. doi: 10.3389/fgene.2020.586462

Duan, P., Xu, J., Zeng, D., Zhang, B., Geng, M., Zhang, G., et al. (2017). Natural variation in the promoter of *GSE5* contributes to grain size diversity in rice. *Mol. Plant* 10 (5), 685–694. doi: 10.1016/j.molp.2017.03.009

Earl, D. A., and VonHoldt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* 14 (8), 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Fitzgerald, M. A., McCouch, S. R., and Hall, R. D. (2009). Not just a grain of rice: the quest for quality. *Trends Plant Sci.* 14 (3), 133–139. doi: 10.1016/j.tplants.2008.12.004

Fujita, N., Miura, S., and Crofts, N. (2022). Effects of various allelic combinations of starch biosynthetic genes on the properties of endosperm starch in rice. *Rice* 15 (1), 1–13. doi: 10.1186/s12284-022-00570-8

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *science* 296 (5576), 2225–2229. doi: 10.1126/science.1069424

Gao, Z., Qian, Q., Liu, X., Yan, M., Feng, Q., Dong, G., et al. (2009). Dwarf 88, a novel putative esterase gene affecting architecture of rice plant. *Plant Mol. Biol.* 71, 265–276. doi: 10.1007/s11103-009-9522-x

Gawenda, I., Thorwarth, P., Günther, T., Ordon, F., and Schmid, K. J. (2015). Genome-wide association studies in elite varieties of German winter barley using single-marker and haplotype-based methods. *Plant Breed.* 134 (1), 28–39. doi: 10.1111/pbr.12237

Guo, L., Chen, W., Tao, L., Hu, B., Qu, G., Tu, B., et al. (2020). *GWC1* is essential for high grain quality in rice. *Plant Sci.* 296, 110497. doi: 10.1016/j.plantsci.2020.110497

Guo, T., Liu, X., Wan, X., Weng, J., Liu, S., Liu, X., et al. (2011). Identification of a stable quantitative trait locus for percentage grains with white chalkiness in rice (*Oryza sativa*). *J. Integr. Plant Biol.* 53 (8), 598–607. doi: 10.1111/j.1744-7909.2011.01041.x

Gupta, A., Bhardwaj, M., and Tran, L. S. P. (2022). Integration of auxin, brassinosteroid and cytokinin in the regulation of rice yield. *Plant Cell Physiol.* 63 (12), 1848–1856. doi: 10.1093/pcp/pcac149

Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. *Adv Genet.* 104, 75–154.

He, Q., and Park, Y. J. (2015). Discovery of a novel fragrant allele and development of functional markers for fragrance in rice. *Mol. Breed.* 35, 1–10. doi: 10.1007/s11032-015-0412-4

Hori, K., Nonoue, Y., Ono, N., Shibaya, T., Ebana, K., Matsubara, K., et al. (2015). Genetic architecture of variation in heading date among Asian rice accessions. *BMC Plant Biol.* 15, 1–16. doi: 10.1186/s12870-015-0501-x

Hori, K., and Sun, J. (2022). Rice grain size and quality. *Rice* 15 (1), 33. doi: 10.1186/s12284-022-00579-z

Hori, K., Suzuki, K., Ishikawa, H., Nonoue, Y., Nagata, K., Fukuoka, S., et al. (2021). Genomic regions involved in differences in eating and cooking quality other than Wx and Alk genes between indica and japonica rice cultivars. *Rice* 14, 1–16. doi: 10.1186/s12284-020-00447-8

Hu, C., Jiang, J., Li, Y., Song, S., Zou, Y., Jing, C., et al. (2022). QTL mapping and identification of candidate genes using a genome-wide association study for heat tolerance at anthesis in rice (*Oryza sativa* L.). *Front. Genet.* 13, 983525. doi: 10.3389/fgene.2022.983525

Hu, J., Wang, Y., Fang, Y., Zeng, L., Xu, J., Yu, H., et al. (2015). A rare allele of *GS2* enhances grain size and grain yield in rice. *Mol. Plant* 8 (10), 1455–1465. doi: 10.1016/j.molp.2015.07.002

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42 (11), 961–967. doi: 10.1038/ng.695

Imran, M., Shafiq, S., Ilahi, S., Ghahramani, A., Bao, G., Dessoky, E. S., et al. (2022). Post-transcriptional regulation of 2-acetyl-1-pyrroline (2-AP) biosynthesis pathway, silicon, and heavy metal transporters in response to Zn in fragrant rice. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.948884

Kikuchi, S., Bheemanahalli, R., Jagadish, K. S., Kumagai, E., Masuya, Y., Kuroda, E., et al. (2017). Genome-wide association mapping for phenotypic plasticity in rice. *Plant Cell Environ.* 40 (8), 1565–1575. doi: 10.1111/pce.12955

Li, P., Chen, Y. H., Lu, J., Zhang, C. Q., Liu, Q. Q., and Li, Q. F. (2022). Genes and their molecular functions determining seed structure, components, and quality of rice. *Rice* 15 (1), 1–27. doi: 10.1186/s12284-022-00562-8

Li, Y., Fan, C., Xing, Y., Yun, P., Luo, L., Yan, B., et al. (2014a). Chalk5 encodes a vacuolar H+-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat. Genet.* 46 (4), 398–404. doi: 10.1038/ng.2923

Li, F., Liu, W., Tang, J., Chen, J., Tong, H., Hu, B., et al. (2010). Rice *DENSE AND ERECT PANICLE 2* is essential for determining panicle outgrowth and elongation. *Cell Res.* 20 (7), 838–849. doi: 10.1038/cr.2010.69

Li, Z. F., Wan, J. M., Xia, J. F., and Zhai, H. Q. (2003). Mapping quantitative trait loci underlying appearance quality of rice grains (*Oryza sativa* L.). *Yi Chuan xue bao= Acta Genetica Sin.* 30 (3), 251–259.

Li, J. Y., Wang, J., and Zeigler, R. S. (2014b). The 3,000 rice genomes project: new opportunities and challenges for future rice research. *Gigascience* 3 (1), 2047–217X. doi: 10.1186/2047-217X-3-8

Liang, R., Qin, R., Yang, C., Zeng, D., Jin, X., and Shi, C. (2019). Identification and characterization of a novel strigolactone-insensitive mutant, Dwarfism with high tillering ability 34 (dhta-34) in rice (Oryza sativa L.). *Biochem. Genet.* 57, 403–420. doi: 10.1007/s10528-018-9896-z

Lin, Y., Zhao, Z., Zhou, S., Liu, L., Kong, W., Chen, H., et al. (2017). *Top Bending Panicle1* is involved in brassinosteroid signaling and regulates the plant architecture in rice. *Plant Physiol. Biochem.* 121, 1–13. doi: 10.1016/j.plaphy.2017.10.001

Lipka, A. E., Kandianis, C. B., Hudson, M. E., Yu, J., Drnevich, J., Bradbury, P. J., et al. (2015). From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24, 110–118. doi: 10.1016/j.pbi.2015.02.010

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28 (18), 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, S., Hua, L., Dong, S., Chen, H., Zhu, X., Jiang, J. E., et al. (2015). *Os MAPK 6*, a mitogen-activated protein kinase, influences rice grain size and biomass production. *Plant J.* 84 (4), 672–681. doi: 10.1111/tpj.13025

Liu, W., Wu, C., Fu, Y., Hu, G., Si, H., Zhu, L., et al. (2009). Identification and characterization of *HTD2*: a novel gene negatively regulating tiller bud outgrowth in rice. *Planta* 230, 649–658. doi: 10.1007/s00425-009-0975-6

Lu, Y., Shah, T., Hao, Z., Taba, S., Zhang, S., Gao, S., et al. (2011). Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. *PloS One* 6 (9), e24861. doi: 10.1371/journal.pone.0024861

Meng, B., Wang, T., Luo, Y., Guo, Y., Xu, D., Liu, C., et al. (2022). Identification and allele combination analysis of rice grain shape-related genes by genome-wide association study. *Int. J. Mol. Sci.* 23 (3), 1065. doi: 10.3390/ijms23031065

Misra, G., Anacleto, R., Badoni, S., Butardo, V.Jr., Molina, L., Graner, A., et al. (2019). Dissecting the genome-wide genetic variants of milling and appearance quality traits in rice. *J. Exp. Bot.* 70 (19), 5115–5130. doi: 10.1093/jxb/erz256

Miura, S., Narita, M., Crofts, N., Itoh, Y., Hosaka, Y., Oitome, N. F., et al. (2022). Improving agricultural traits while maintaining high resistant starch content in rice. *Rice* 15 (1), 1–16. doi: 10.1186/s12284-022-00573-5

Mo, Z., Huang, J., Xiao, D., Ashraf, U., Duan, M., Pan, S., et al. (2016). Supplementation of 2-Ap, Zn and La improves 2-acetyl-1-pyrroline concentrations in detached aromatic rice panicles in *vitro*. *PloS One* 11 (2), e0149523. doi: 10.1371/journal.pone.0149523

N'Diaye, A., Haile, J. K., Cory, A. T., Clarke, F. R., Clarke, J. M., Knox, R. E., et al. (2017). Single marker and haplotype-based association analysis of semolina and pasta colour in elite durum wheat breeding lines using a high-density consensus map. *PloS One* 12 (1), e0170941. doi: 10.1371/journal.pone.0187178

Nguyen, K. M., and Busino, L. (2020). The biology of F-box proteins: the SCF family of E3 ubiquitin ligases. *Adv Exp Med Biol* 67, 53–60. doi: 10.1007/978-981-15-1025-0_8

Nirmaladevi, G., Padmavathi, G., Kota, S., and Babu, V. R. (2015). Genetic variability, heritability and correlation coefficients of grain quality characters in rice (Oryza sativa L.). *SABRAO J. Breed. Genet.* 47 (4), 424–433.

Poonlaphdecha, J., Gantet, P., Maraval, I., Sauvage, F. X., Menut, C., Morère, A., et al. (2016). Biosynthesis of 2-acetyl-1-pyrroline in rice calli cultures: Demonstration of 1-pyrroline as a limiting substrate. *Food Chem.* 197, 965–971. doi: 10.1016/j.foodchem.2015.11.060

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38 (8), 904–909. doi: 10.1038/ng1847

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155 (2), 945–959. doi: 10.1093/genetics/155.2.945

Qiu, X., Yang, J., Zhang, F., Niu, Y., Zhao, X., Shen, C., et al. (2021). Genetic dissection of rice appearance quality and cooked rice elongation by genome-wide association study. *Crop J.* 9 (6), 1470–1480. doi: 10.1016/j.cj.2020.12.010

Quero, G., Gutiérrez, L., Monteverde, E., Blanco, P., Perez de Vida, F., Rosas, J., et al. (2018). Genome-wide association study using historical breeding populations discovers genomic regions involved in high-quality rice. *Plant Genome* 11 (3), 170076. doi: 10.3835/plantgenome2017.08.0076

Reig-Valiente, J. L., Marqués, L., Talón, M., and Domingo, C. (2018). Genome-wide association study of agronomic traits in rice cultivated in temperate regions. *BMC Genomics* 19, 706. doi: 10.1186/s12864-018-5086-y

Ren, W. L., Wen, Y. J., Dunwell, J. M., and Zhang, Y. M. (2018). pKWmEB: integration of Kruskal–Wallis test with empirical Bayes under polygenic background control for multi-locus genome-wide association study. *Heredity* 120 (3), 208–218. doi: 10.1038/s41437-017-0007-4

Sahu, P. K., Sharma, D., Mondal, S., Kumar, V., Singh, S., Baghel, S., et al. (2017). Genetic variability for grain quality traits in indigenous rice landraces of Chhattisgarh India. *J. Exp. Biol. Agric. Sci.* 5 (4), 439–455. doi: 10.18006/2017.5(4).439.455

Sanchez, D. L., Samonte, S. O., and Wilson, L. T. (2023). Genetic architecture of head rice and rice chalky grain percentages using genome-wide association studies. *Front. Plant Sci.* 14, 1274823. doi: 10.3389/fpls.2023.1274823

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.1038/ng.2314

Selvaraj, R., Singh, A. K., Singh, V. K., Abbai, R., Habde, S. V., Singh, U. M., et al. (2021). Superior haplotypes towards development of low glycemic index rice with preferred grain and cooking quality. *Sci. Rep.* 11 (1), 1–15. doi: 10.1038/s41598-021-87964-8

Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q., et al. (2016). *OsSPL13* controls grain size in cultivated rice. *Nat. Genet.* 48 (4), 447–456. doi: 10.1038/ng.3518

Song, X. J., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39 (5), 623–630. doi: 10.1038/ng2014

Sood, B. C., and Siddiq, E. A. (1978). A rapid technique for scent determination in rice [India]. *Indian J. Genet. Plant Breeding* 38, 268–271.

Sun, S., Wang, L., Mao, H., Shao, L., Li, X., Xiao, J., et al. (2018). A G-protein pathway determines grain size in rice. *Nat. Commun.* 9 (1), 851. doi: 10.1038/s41467-018-03141-y

Tamba, C. L., Ni, Y. L., and Zhang, Y. M. (2017). Iterative sure independence screening EM-Bayesian LASSO algorithm for multi-locus genome-wide association studies. *PloS Comput. Biol.* 13 (1), e1005357. doi: 10.1371/journal.pcbi.1005357

Tamba, C. L., and Zhang, Y. M. (2018). A fast mrMLM algorithm for multi-locus genome-wide association studies. *biorxiv* 10, 341784. doi: 10.1101/341784

Tan, Y. F., Sun, M., Xing, Y. Z., Hua, J. P., Sun, X. L., Zhang, Q. F., et al. (2001). Mapping quantitative trait loci for milling quality, protein content and color characteristics of rice using a recombinant inbred line population derived from an elite rice hybrid. *Theor. Appl. Genet.* 103, 1037–1045. doi: 10.1007/s001220100665

Tibbs Cortes, L., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14 (1), e20077. doi: 10.1002/tpg2.20077

Varshney, R. K., Terauchi, R., and McCouch, S. R. (2014). Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. *PloS Biol.* 12 (6), e1001883. doi: 10.1371/journal.pbio.1001883

Vemireddy, L. R., Noor, S., Satyavathi, V. V., Srividhya, A., Kaliappan, A., Parimala, S. R. N., et al. (2015). Discovery and mapping of genomic regions governing economically important traits of Basmati rice. *BMC Plant Biol.* 15, 1–19. doi: 10.1186/s12870-015-0575-5

Verma, R. K., Chetia, S. K., Dey, P. C., Rahman, A., Saikia, S., Sharma, V., et al. (2021). Genome-wide association studies for agronomical traits in winter rice accessions of Assam. *Genomics* 113 (3), 1037–1047. doi: 10.1016/j.ygeno.2020.11.033

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101 (1), 5–22. doi: 10.1016/j.ajhg.2017.06.005

Voorman, A., Lumley, T., McKnight, B., and Rice, K. (2011). Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PloS One* 6 (5), e19416. doi: 10.1371/journal.pone.0019416

Wakte, K., Zanan, R., Hinge, V., Khandagale, K., Nadaf, A., and Henry, R. (2017). Thirty-three years of 2-acetyl-1-pyrroline, a principal basmati aroma compound in scented rice (*Oryza sativa* L.): a status review. *J. Sci. Food Agric.* 97 (2), 384–395. doi: 10.1002/jsfa.7875

Wan, X. Y., Wan, J. M., Weng, J. F., Jiang, L., Bi, J. C., Wang, C. M., et al. (2005). Stability of QTLs for rice grain dimension and endosperm chalkiness characteristics across eight environments. *Theor. Appl. Genet.* 110, 1334–1346. doi: 10.1007/s00122-005-1976-x

Wang, D. R., Agosto-Pérez, F. J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., et al. (2018). An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.* 9 (1), 3519. doi: 10.1038/s41467-018-05538-1

Wang, J., Chen, Z., Zhang, Q., Meng, S., and Wei, C. (2020). The NAC transcription factors *OsNAC20* and *OsNAC26* regulate starch and storage protein synthesis. *Plant Physiol.* 184 (4), 1775–1791. doi: 10.1104/pp.20.00984

Wang, S. B., Feng, J. Y., Ren, W. L., Huang, B., Zhou, L., Wen, Y. J., et al. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Sci. Rep.* 6, 19444. doi: 10.1038/srep19444

Wang, S., Li, S., Liu, Q., Wu, K., Zhang, J., Wang, S., et al. (2015). The *OsSPL16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* 47 (8), 949–954. doi: 10.1038/ng.3352

Wang, X., Pang, Y., Wang, C., Chen, K., Zhu, Y., Shen, C., et al. (2017). New candidate genes affecting rice grain appearance and milling quality detected by genome-wide and gene-based association analyses. *Front. Plant Sci.* 7, 1998. doi: 10.3389/fpls.2016.01998

Waugh, R., Thomas, B., Flavell, A., Ramsay, L., Comadran, J., and Russell, J. (2014). Genome-wide association scans (GWAS). *Biotechnol. Approaches to Barley Improvement* 69, 345–365. doi: 10.1007/978-3-662-44406-1_18

Wen, Y. J., Zhang, H., Ni, Y. L., Huang, B., Zhang, J., Feng, J. Y., et al. (2018). Methodological implementation of mixed linear models in multi-locus genome-wide association studies. *Briefings Bioinf.* 19 (4), 700–712. doi: 10.1093/bib/bbw145

Xie, L., Tang, S., Chen, N., Luo, J., Jiao, G., Shao, G., et al. (2013). Rice grain morphological characteristics correlate with grain weight and milling quality. *Cereal Chem.* 90 (6), 587–593. doi: 10.1094/CCHEM-03-13-0055-R

Xing, Y., and Zhang, Q. (2010). Genetic and molecular bases of rice yield. *Annu. Rev. Plant Biol.* 61, 421–442. doi: 10.1146/annurev-arplant-042809-112209

Xu, C., Liu, Y., Li, Y., Xu, X., Xu, C., Li, X., et al. (2015). Differential expression of *GS5* regulates grain size in rice. *J. Exp. Bot.* 66 (9), 2611–2623. doi: 10.1093/jxb/erv058

Yan, P., Zhu, Y., Wang, Y., Ma, F., Lan, D., Niu, F., et al. (2022). A new RING finger protein, *PLANT ARCHITECTURE and GRAIN NUMBER 1*, affects plant architecture and grain yield in rice. *Int. J. Mol. Sci.* 23 (2), 824. doi: 10.3390/ijms23020824

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48 (8), 927–934. doi: 10.1038/ng.3596

Yu, S., Jiang, S., Zhu, L., Zhang, J., and Jin, Q. (2013). Effects of acoustic frequency technology on rice growth, yield and quality. *Trans. Chin. Soc. Agric. Eng.* 29 (2), 141–147.

Yuan, H., Fan, S., Huang, J., Zhan, S., Wang, S., Gao, P., et al. (2017). *08SG2/OsBAK1* regulates grain size and number, and functions differently in Indica and Japonica backgrounds in rice. *Rice* 10 (1), 1–12. doi: 10.1186/s12284-017-0165-2

Zhang, J., Feng, J. Y., Ni, Y. L., Wen, Y. J., Niu, Y., Tamba, C. L., et al. (2017). pLARmEB: integration of least angle regression with empirical Bayes for multilocus genome-wide association studies. *Heredity* 118 (6), 517–524. doi: 10.1038/hdy.2017.8

Zhang, F., Wang, Y., and Deng, H. W. (2008). Comparison of population-based association study methods correcting for population stratification. *PloS One* 3 (10), e339. doi: 10.1371/journal.pone.0003392

Zhang, P., Zhong, K., Zhong, Z., and Tong, H. (2019). Genome-wide association study of important agronomic traits within a core collection of rice (*Oryza sativa* L.). *BMC Plant Biol.* 19, 1–12. doi: 10.1186/s12870-019-1842-7

Zhang, H., Zhou, L., Xu, H., Wang, L., Liu, H., Zhang, C., et al. (2019). The *qSAC3* locus from indica rice effectively increases content under a variety of conditions. *BMC Plant Biol.* 19, 1–11. doi: 10.1186/s12870-019-1860-5

Zhang, C., Zhu, J., Chen, S., Fan, X., Li, Q., Lu, Y., et al. (2019). *Wxlv*, the ancestral allele of rice Waxy gene. *Mol. Plant* 12 (8), 1157–1166. doi: 10.1016/j.molp.2019.05.011

Zhao, X., Daygon, V. D., McNally, K. L., Hamilton, R. S., Xie, F., Reinke, R. F., et al. (2016). Identification of stable QTLs causing chalk in rice grains in nine environments. *Theor. Appl. Genet.* 129, 141–153. doi: 10.1007/s00122-015-2616-8

Zhao, D. S., Li, Q. F., Zhang, C. Q., Zhang, C., Yang, Q. Q., Pan, L. X., et al. (2018). GS9 acts as a transcriptional activator to regulate rice grain shape and appearance quality. *Nat. Commun.* 9 (1), 1240. doi: 10.1038/s41467-018-03616-y

Zhong, H., Liu, S., Sun, T., Kong, W., Deng, X., Peng, Z., et al. (2021). Multi-locus genome-wide association studies for five yield-related traits in rice. *BMC Plant Biol.* 21 (1), 364. doi: 10.1186/s12870-021-03146-8

Zhou, L., Chen, L., Jiang, L., Zhang, W., Liu, L., Liu, X., et al. (2009). Fine mapping of the grain chalkiness QTL *qPGWC-7* in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 118, 581–590. doi: 10.1007/s00122-008-0922-0

Zhou, H., Xia, D., and He, Y. (2020). Rice grain quality—traditional traits for high quality rice and health-plus substances. *Mol. Breed.* 40, 1–17. doi: 10.1007/s11032-019-1080-6

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 05 March 2024
DOI 10.3389/fpls.2024.1337463

OPEN ACCESS

EDITED BY
Ting Peng,
Henan Agricultural University, China

REVIEWED BY
Zenglu Li,
University of Georgia, United States
Anilkumar C,
National Rice Research Institute (ICAR), India

*CORRESPONDENCE
Siwaret Arikit
siwaret.a@ku.th
Vinitchan Ruanjaichon
vinitchan.rua@biotec.or.th

†These authors have contributed equally to this work

RECEIVED 13 November 2023
ACCEPTED 19 February 2024
PUBLISHED 05 March 2024

CITATION
Khammona K, Dermail A, Suriharn K, Lübberstedt T, Wanchana S, Thunnom B, Poncheewin W, Toojinda T, Ruanjaichon V and Arikit S (2024) Accelerating haploid induction rate and haploid validation through marker-assisted selection for *qhir1* and *qhir8* in maize. *Front. Plant Sci.* 15:1337463. doi: 10.3389/fpls.2024.1337463

Accelerating haploid induction rate and haploid validation through marker-assisted selection for *qhir1* and *qhir8* in maize

Kanogporn Khammona[1†], Abil Dermail[2†], Khundej Suriharn[2,3], Thomas Lübberstedt[4], Samart Wanchana[5], Burin Thunnom[5], Wasin Poncheewin[5], Theerayut Toojinda[5], Vinitchan Ruanjaichon[5*] and Siwaret Arikit[1,6*]

[1]Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University, Nakhon Pathom, Thailand, [2]Department of Agronomy, Faculty of Agriculture, Khon Kaen University, Khon Kaen, Thailand, [3]Plant Breeding Research Center for Sustainable Agriculture, Faculty of Agriculture, Khon Kaen University, Khon Kaen, Thailand, [4]Department of Agronomy, Iowa State University, Ames, IA, United States, [5]National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand, [6]Rice Science Center, Kasetsart University, Nakhon Pathom, Thailand

Doubled haploid (DH) technology becomes more routinely applied in maize hybrid breeding. However, some issues in haploid induction and identification persist, requiring resolution to optimize DH production. Our objective was to implement simultaneous marker-assisted selection (MAS) for *qhir1* (*MTL/ZmPLA1/NLD*) and *qhir8* (*ZmDMP*) using TaqMan assay in F_2 generation of four BHI306-derived tropical × temperate inducer families. We also aimed to assess their haploid induction rate (HIR) in the F_3 generation as a phenotypic response to MAS. We highlighted remarkable increases in HIR of each inducer family. Genotypes carrying *qhir1* and *qhir8* exhibited 1 – 3-fold higher haploid frequency than those carrying only *qhir1*. Additionally, the *qhir1* marker was employed for verifying putative haploid seedlings at 7 days after planting. Flow cytometric analysis served as the gold standard test to assess the accuracy of the *R1-nj* and the *qhir1* marker. The *qhir1* marker showed high accuracy and may be integrated in multiple haploid identifications at early seedling stage succeeding pre-haploid sorting via *R1-nj* marker.

KEYWORDS

hybrid breeding, doubled haploid, haploid induction, haploid identification, molecular assay

Introduction

Maize is one of the most important cereal crops in the world as food, feed, and fuel (Prasanna, 2012). The success of hybrid maize breeding relies on robust pipelines of germplasm, genetics, phenotyping, and selection processes (Cooper et al., 2014). Traditionally, the breeding process for the market release of a new cultivar extended over a decade, until the advent of doubled haploid (DH) technology (Chaikam et al., 2019). A notable advantage of DH technology is associated with the substantial reduction of breeding cycles required to develop fully homozygous lines within just two generations (Geiger and Gordillo, 2009). Haploids can be produced *in vitro* or *in vivo*. The *in vitro* method requires laboratory procedures, where gametophytic tissues such as microspores and egg cells are used to produce paternal and maternal haploids, respectively. However, this method gains low success rates due to the high levels of genotype dependency (Jacquier et al., 2021). The *in vivo* method involves four main steps: (1) haploid induction, (2) haploid identification, (3) haploid genome doubling, and (4) self-pollination of haploid plants to obtain DH_0 seeds (Chaikam et al., 2019). For maternal haploid induction, haploid inducers are used as male parents to pollinate source germplasm for haploid induction. Efficient DH line production depends on the availability of inducer genotypes with high induction ability.

In 2012, a QTL study involving four populations, all sharing the inbred inducer UH400 as common parent, identified 8 QTL. Notably, *qhir1* and *qhir8* emerged as two major QTL located on chromosomes 1 and 9, explaining 66% and 20% of the genetic variance, respectively (Prigge et al., 2012). The *qhir1* region in bin 1.04 plays pivotal roles in triggering haploid induction, gametophytic segregation distortion, and embryo abortion (Barret et al., 2008; Prigge et al., 2012; Xu et al., 2013). Mutation of the gene *MTL/ZmPLA1/NLD* in *qhir1* has been shown to generate an average haploid induction rate (HIR) up to 6.7% (Gilles et al., 2017; Kelliher et al., 2017; Liu et al., 2017). However, *qhir1* is not sufficient for commercial productions of DH lines. To fully leverage this technology, the average HIR of modern haploid inducers should surpass 10% (Hu et al., 2016). Zhong et al. (2019) discovered a novel gene named *ZmDMP* underlying QTL *qhir8*. A mutation of *ZmDMP* markedly enhances haploid induction, resulting in a 2–3-fold increase in HIR. It is important to note that both *MTL/ZmPLA1/NLD* and *ZmDMP* act synergistically, suggesting the potential for a substantial 5–6-fold increase in the HIR when both mutations are present (Zhong et al., 2019). Marker assisted selection (MAS) for *qhir1* has been applied to improve the HIR of maternal haploid inducers in different maize backgrounds. For instance, Chaikam et al. (2018) were able to obtain promising second-generation Tropically Adapted Inducer Lines (2GTAILs) with an average HIR of 13.1%, a 48.9% improvement over TAILs. Liu et al. (2022) developed an elite oil haploid inducer, CHOI4, with an averaged HIR of 15.8%, a 58.0% increase compared to CAU2, the founder parent of CHOI4. While these results are promising, further enhancements could be achieved through MAS for two loci, *qhir1* and *qhir8*. Considering that HIR is a polygenic trait, selection of a single locus may not be sufficient to obtain inducers with optimum HIR (Dong et al., 2014). Nevertheless, limited evidence exists to illustrate the feasibility of MAS for both loci simultaneously in breeding haploid inducers for high HIR.

Haploids are commonly identified via the *R1-navajo* (*R1-nj*), a dominant monogenic biomarker (Nanda and Chase, 1966) integrated in haploid inducers. This marker distinguishes progeny seeds derived from haploid induction based on anthocyanin expression in different parts of the kernel. Haploid kernels show a purple crown in the endosperm but a colorless scutellum in the embryo, while diploids express both purple endosperm and embryo (Dermail et al., 2021). Despite practical and non-destructive features, the effectiveness of *R1-nj* expressions may be constrained by the presence of dominant *C1* anthocyanin inhibitors (Chaikam et al., 2015), naturally occurring anthocyanins in donor germplasm (Chaikam et al., 2016), morpho-physiological kernel properties (Prigge et al., 2011; Trentin et al., 2022), and environments (Sintanaparadee et al., 2022; Dermail et al., 2023; Thawarorit et al., 2023). These factors contribute to high misclassification rates (MCRs), hindering selection gains on HIR and emphasizing the need for alternative markers for haploid selection. While simple sequence repeat (SSR) has been successfully used in maize haploid identification (Qiu et al., 2014; Dong et al., 2018; Li et al., 2021), the practical use of SNP markers for that purpose is still lacking. Since most paternal chromosomes of inducers are excluded from haploid embryonic cells within a week after pollination, the haploid individuals carry only maternal chromosomes from the donor germplasm (Zhao et al., 2013). Codominant SNP markers can differentiate between homozygotes (donor female) and heterozygotes (F_1 diploids). Considering remarkable allelic variation for *qhir1* and *qhir8* haploid inducers versus non-inducer genotypes, there is an encouraging prospect of applying these loci for haploid identification using TaqMan probes. Kelliher et al. (2017) employed TaqMan assays for *qhir1*, proposing that haploids carry zero copies of the *mtl* allele and two copies of the maternal *MTL* allele, whereas diploids carry one copy of the *mtl* allele and one copy of the *MTL* allele.

Our study aimed to utilize the *qhir1* and *qhir8* loci in marker-assisted selection, with a dual focus on breeding haploid inducers for high HIR and accurately identifying true haploids in maize. We hypothesized that (i) inducer genotypes carrying *qhir1* and *qhir8* should demonstrate a higher capacity to induce haploids compared to those carrying *qhir1* alone and (ii) molecular markers using TaqMan assay are more reliable than the *R1-nj* marker when validated with flow cytometry. This study will provide an insight into the advantages of molecular assays, especially TaqMan probes, to accelerate the improvements of haploid inducers underpinning HIR. Additionally, it seeks to enhance the accuracy of identifying true haploids at early seedling stage.

Materials and methods

Breeding scheme, haploid induction, and HIR evaluation

A temperate inbred inducer, BHI306, and four tropical inducer families (K8, K11, KHI49, and KHI54) were selected as founder parents. The BHI306 genotype, an RWS/RWK-76-derived haploid

inducer, has 10–15% of HIR, carries both *qhir1* and *qhir8* loci (Supplementary Table S1 and Supplementary Figure S1), kernel anthocyanin *R1-nj* and red root *Pl-1* selectable markers. BHI306 was developed by the DH Facility of Iowa State University (DHF-ISU) (https://www.doubledhaploid.biotech.iastate.edu/). Four genotypes, K8, K11, KHI49, and KHI54 belong to qhir1−/qhir8− group (Supplementary Table S1 and Supplementary Figure S1), Stock-6-derived haploid inducers, had low HIRs (<6.0%) but possess favorable tropical adaptations. These genotypes were developed by the Plant Breeding Research Center for Sustainable Agriculture of Khon Kaen University in Thailand (Dermail et al., 2021; Sintanaparadee et al., 2022; Thawarorit et al., 2023). A 1 × 4 factorial mating scheme was performed by assigning BHI306 as a male and four tropical inducers as females to establish four tropical × temperate inducer base populations including K8/BHI306, K11/BHI306, KHI49/BHI306, and KHI54/BHI306. In the F_2 generation, approximately 100 F_2 seedlings per inducer population underwent random marker-assisted selection (MAS) for *qhir1* and *qhir8*. Plants carrying *qhir1* only and both *qhir1* and *qhir8* were labeled as *qhir1+/qhir8−* and *qhir1+/qhir8+* genotypes, respectively. These targeted genotypes were subsequently transplanted into the field and self-pollinated to obtain F_3 seeds. At that generation, we did not perform haploid induction. Thus, there was no preliminary information regarding the actual HIR. At the F_3 generation, repeated genotyping of each individual plant and phenotyping on actual HIR were performed in each population (Supplementary Table S1).

Maternal haploid induction was performed to evaluate HIRs. A commercial hybrid Pacific789 (P789), developed by Pacific Seeds, Thailand, was used as a donor female. This genotype is resistant to tropical diseases, high-yielding, and large-seeded with flat embryos, facilitating haploid selection based on the *R1-nj* marker at the seed stage. Each F_3 inducer plant in each *qhir* genotype and family was used to pollinate four donor ears to minimize the errors due to unstable inducer pollen. Shoot bagging and detasseling of donor plants were routinely performed to prevent pollen contamination.

Haploid seed was selected via the *R1-nj* marker at the seed stage. Haploids showed a purple crown endosperm but a colorless embryo, while diploids expressed purple colorations on both crown endosperm and embryo (Nanda and Chase, 1966; Dermail et al., 2023). The HIR was calculated as the frequency of haploid seeds per induction cross, as follows:

$$\text{HIR}\ (\%) = \frac{\text{seed number of putative haploid}}{\text{seed set}} \text{ x } 100$$

where seed set represents the total seed number of haploid seeds, diploid seeds, and the seeds without the *R1-nj* marker.

About 10 putative haploid seeds per genotype in each inducer family were sampled for further true haploid confirmation through molecular assays.

Marker development

Two TaqMan® markers (*qhir1* and *qhir8*) for two targeted genes namely *MATRILINEAL* (*MTL/ZmPLA1/NLD*) and *ZmDMP*, respectively, were constructed (Figure 1). The marker for the *MTL* gene (GRMZM2G471240) was developed at 4 bp (CGAG) insertion in the 4th exon of the gene that led to premature stop codon (Gilles et al., 2017; Kelliher et al., 2017; Liu et al., 2017). The *ZmDMP* gene (GRMZM2G465053) was developed at single nucleotide substitution from T to C at 131 bp on coding sequence that led to amino acid change from methionine to threonine (Zhong et al., 2019).

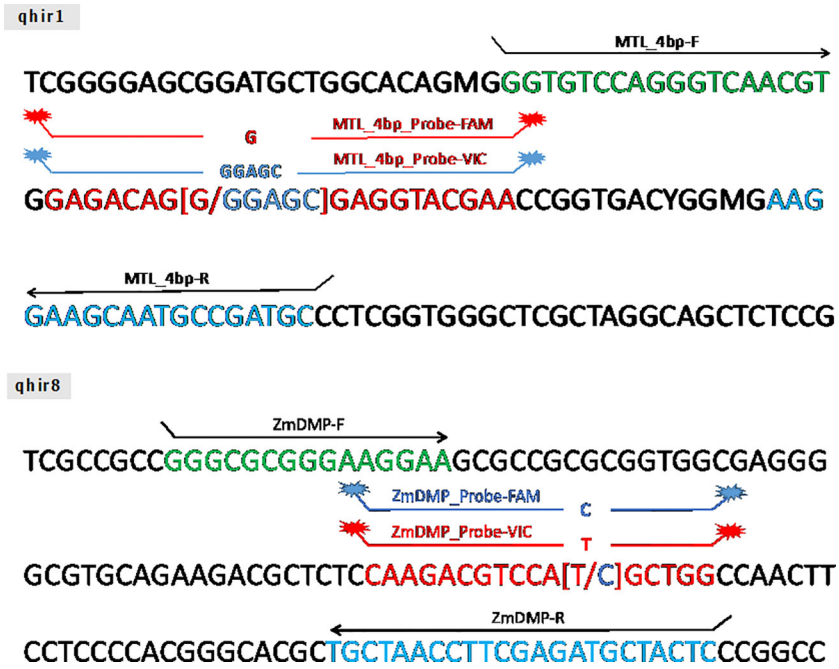


FIGURE 1
The schematic of TaqMan® probe design on *MATRILINEAL* gene (GRMZM2G471240) and *ZmDMP* gene (GRMZM2G465053).

Genotyping and DNA extraction

High-quality genomic DNA (gDNA) was isolated from maize leaves at 14 days after germination using the DNeasy® Plant Mini Kit (QIAGEN, Germany). Genotyping for *qhir1* and *qhir8* markers was carried out with ready-to-order TaqMan assays (Thermo Fisher Scientific, Watham, MA USA) (Figure 1). In the amplification process, 20 ng gDNA was utilized. For the PCR reaction, the total volume was 5 µl composed of 2 µl of template DNA, 1.5 µl of 2X TaqMan® Gene Expression Master Mix (Thermo Fisher Scientific, Watham, MA USA), 0.0375 µl of TaqMan assay, and 1.4625 µl of dH2O. The PCR cycling conditions were set at 95°C for 5 min, followed by 36 cycles at 94°C for 30 s, 60°C for 1 min, and 60°C for 2 min. For the PCR product, the amplicons were melted at 60°C using QuantStudio 6 Real-Time PCR Systems (Thermo Fisher Scientific, Watham, MA USA) for 30 s to detect single nucleotide polymorphism (SNP).

Flow cytometry analysis

Three subsets of populations derived from induction crosses between female donor P789 and three male inducers, BHI306, KHI49/BHI306, and KHI54/BHI306, were used for haploid validation via flow cytometry analysis. The number of samples was 24, derived from false positives previously assumed as putative haploids based on the *R1-nj* marker but eventually true diploids regarding the *qhir1* marker. Those 24 samples composed of 1 putative haploid of P789/BHI306, 10 putative haploids of P789/(KHI49/BHI306), and 13 putative haploids of P789/(KHI54/BHI306). The FC analysis on those 24 samples served as the gold standard classification method to verify if *qhir1* marker is reliable to determine the true haploids. The FC graph of each sample can be found in the Supplementary Figure S3, and the result of FC analysis corresponding to the *qhir1* marker assay can be found in Table 1.

Two maize leaves at 14 days after germination were cut about 3 cm in length (50-100 mg fresh weight) and placed into a plastic petri dish on ice. Then, 1.5 ml of LB01 buffer (15 mM Tris, 2 mM Na2EDTA, 0.5 mM spermine.4HCl, 80 mM KCl, 20 mM NaCl, 0.1% (v/v) Triton X-100, pH 7.5) (Doležel, 1997) was added, and the leaves were chopped in this buffer using a razor blade to facilitate the release of the nuclei (Pfosser et al., 1995). After that, 500 ul of the cell solution was transferred into a 1.5 ml tube. Propidium Iodide 1 mg/mlI-stained nuclei and RNaseA were then added to the solution. The BD Accuri™ C6 Plus flow cytometer (BD Biosciences, USA) was employed for measurement. The ploidy status of each sample can be determined by the fluorescence intensity of stained cell nuclei isolated from plant tissue. The peak value (G1) of haploid is commonly set to half of the diploid reference (Supplementary Figure S3).

Statistical analysis

A total of 237 inducer plants were evaluated for HIR performance including K8/BHI306 (32 plants), K11/BHI306 (54 plants), KHI49/BHI306 (52 plants), and KHI54/BHI306 (99 plants). Each induced donor ear was represented as a technical replicate, resulting in four replications for each inducer plant within each genotype and family. The HIR for each genotype was calculated as the mean HIR across these four replications. The data were subjected to the unpaired samples t-test with 95% confidence interval (CI), Tukey's Honestly Significant Difference (HSD) Test at 5%, and linear regression analysis.

Results

Haploid inducer breeding via marker-assisted selection for *qhir1* and *qhir8*

The median HIR of *qhir1+/qhir8+* genotypes was significantly ($P<0.01$) higher than that of the *qhir1+/qhir8−* genotypes within each F_3 inducer family (Figure 2A). Across the four families, the average HIR for the *qhir1+/qhir8+* genotype ranged from 3.85 to 9.48%, while the average HIR for the *qhir1+/qhir8−* genotype was significantly ($P<0.01$) lower, ranging from 1.18 to 4.89% (Figure 2B; Table 2). This suggests that inducer genotypes fixed for both targeted loci for HIR, *qhir1* and *qhir8*, have remarkable abilities to induce haploids, showing an increase of 3–5% or 1–3-fold higher than inducer genotypes fixed for *qhir1* only.

The proportion of phenotypic variation explained (PVE) across inducer families ranged from 17% to 39% (Table 2). These values, within acceptable ranges, indicated that MAS for two loci was effective in identifying haploid inducers with high HIR. We also found that the HIR between families within the same *qhir1+/qhir8+* genotype was significantly different (Table 2). For instance, families K8/BHI306 and K11/BHI306 demonstrated a significantly higher HIR than families KHI-49/BHI306 and KHI-54/BHI306. The evidence of low %PVE (<50%) (Table 2), outliers, and overlapping values between two inducer groups on HIR (Figure 2), suggests the potential existence of other minor QTL influencing HIR.

Haploid validation via *qhir1* marker and flow cytometry analysis

Marker-assisted selection (MAS) for *qhir1* was applied to validate putative haploids and diploids derived from the *R1-nj* marker system as a preliminary haploid identification among the F_1 progenies of induction crosses. Both parents, BHI306 and P789, were included as positive and negative controls for *qhir1*, respectively (Figure 3). Through the TaqMan assay, all samples of P789, the female donor, were found to be homozygous for *qhir1−* (G/G), while all samples of BHI306, the male inducer, were homozygous *qhir1+* (GGAGC/GGAGC). The sample progenies were then distributed into two pools according to haplotypes: (1) the diploid class, heterozygous for *qhir1* (G/GGAGC) and (2) the haploid class, homozygous for *qhir1−* (G/G), which was grouped with the donor female P789 (Figure 3B, Table 3). Similar results for other populations can be seen in Supplementary Table S2 and Supplementary Figure S2. A few numbers of false positives were

TABLE 1 Haploid validation via *qhir1* marker and flow cytometry (FC) analysis of 24 false positives derived from subsets of F_1 induction crosses between female donor P789 and three male inducers BHI306, KHI49/BHI306, and KHI54/BHI306.

No.	Sample name	*qhir1*	FC	*qhir1* vs. FC	
				R^2	*p-value*
1	P789/BHI306-F_1_n-5	2n	2n	1.00	2.2E−16
2	P789/(KHI49/BHI-F_3)-F_1_n-1-2-6	2n	2n		
3	P789/(KHI49/BHI-F_3)-F_1_n-1-3-2	2n	2n		
4	P789/(KHI49/BHI-F_3)-F_1_n-1-4-8	2n	2n		
5	P789/(KHI49/BHI-F_3)-F_1_n-1-6-5	2n	2n		
6	P789/(KHI49/BHI-F_3)-F_1_n-1-7-1	2n	2n		
7	P789/(KHI49/BHI-F_3)-F_1_n-1-14-7	2n	2n		
8	P789/(KHI49/BHI-F_3)-F_1_n-1-15-3	2n	2n		
9	P789/(KHI49/BHI-F_3)-F_1_n-1-20-10	2n	2n		
10	P789/(KHI49/BHI-F_3)-F_1_n-2-2-6	2n	2n		
11	P789/(KHI49/BHI-F_3)-F_1_n-2-2-9	2n	2n		
12	P789/(KHI54/BHI-F_3)-F_1_n-1-1-3	2n	2n		
13	P789/(KHI54/BHI-F_3)-F_1_n-1-1-6	2n	2n		
14	P789/(KHI54/BHI-F_3)-F_1_n-1-2-3	2n	2n		
15	P789/(KHI54/BHI-F_3)-F_1_n-1-3-4	2n	2n		
16	P789/(KHI54/BHI-F_3)-F_1_n-1-4-3	2n	2n		
17	P789/(KHI54/BHI-F_3)-F_1_n-1-10-1	2n	2n		
18	P789/(KHI54/BHI-F_3)-F_1_n-1-11-1	2n	2n		
19	P789/(KHI54/BHI-F_3)-F_1_n-1-12-7	2n	2n		
20	P789/(KHI54/BHI-F_3)-F_1_n-1-16-2	2n	2n		
21	P789/(KHI54/BHI-F_3)-F_1_n-1-17-1	2n	2n		
22	P789/(KHI54/BHI-F_3)-F_1_n-1-18-1	2n	2n		
23	P789/(KHI54/BHI-F_3)-F_1_n-1-19-2	2n	2n		
24	P789/(KHI54/BHI-F_3)-F_1_n-2-3-10	2n	2n		

R^2 coefficient of determination.
All 24 false positives were previously classified as putative haploids based on the *R1-nj* marker, but then they were verified as true haploids based on the *qhir1* marker. The result of the FC graph on each of the 24 samples can be found in Supplementary Figure S3.

found in putative haploid populations derived from induction crosses, accounting for 1, 10, and 13 samples in populations P789/BHI306-F_1, P789/(KHI49/BHI306-F_3)-F_1, and P789/(KHI54/BHI306-F_3)-F_1, respectively (Table 3). The reliability of the *qhir1* for haploid determination was further validated by flow cytometric analysis. We found that the result of FC analysis among 24 false positives (Supplementary Figure S3) corresponded to the *qhir1* marker, as indicated by R^2 = 1.00 (Table 1). This implies that the *qhir1* marker using the TaqMan assay was effective to identify true haploids, indicated by a 0-false positive rate, which could thus serve as an alternative gold standard test compared to flow cytometry in future. It also suggests that a single SNP marker, like *qhir1*, is ultimately sufficient for haploid identification to reduce the cost of genotyping.

Discussion

Marker-assisted selection (MAS) may accelerate breeding programs by indirectly selecting target traits using molecular markers tightly linked to underlying genes (Xu and Crouch, 2008). Plant breeders can benefit from this approach especially when targeted traits pose challenges for improvement through traditional phenotypic selections. Technical issues such as resource intensiveness and genetic properties like low heritability, complex inheritance, and presence of recessive alleles make phenotypic selection difficult (Koebner, 2004; Collard et al., 2005; Xu et al., 2005). It is suitable for our breeding objectives to accelerate the rates of haploid induction (HIR) possessing multiple recessive alleles and QTL (Prigge et al., 2012) and prone to the environments of haploid induction (Kebede et al., 2011; De La Fuente et al., 2018; Sintanaparadee et al., 2022).

The effectiveness of MAS for *qhir1* has been reported in the breeding high-oil inducers (Dong et al., 2014) and the development of CIMMYT second-generation Tropically Adapted Inducer Lines (CIM2GTAILs) (Chaikam et al., 2018). Trentin et al. (2020) suggested a stratified MAS approach, initially targeting the *mtl* allele or *qhir1* in a large F_2 population and later for *zmdmp* allele or *qhir8* in F_3 plants carrying the *mtl* allele or *qhir1*. In our study, we validated the efficacy of simultaneous MAS for *qhir1* and *qhir8* in F_2 segregating populations, leading to enhanced HIR in F_3 genotypes by 1–3-fold. We also noticed that the genotype of *qhir1−/qhir8+* and heterozygous *qhir1/qhir8+* showed lower HIR than genotypes with *qhir1+* (data not shown). Our findings align with Zhong et al. (2019), who identified a novel mutation in the *ZmDMP* gene in the CAUHOI (*qhir1+*) genotype and demonstrated its impact on HIR. They found that the genotype with qhir1+/qhir8+ exhibited inflating HIR by 5–6-fold compared to *qhir1+/qhir8−*. The implementation of MAS in the early generations proves beneficial by significantly reducing the number of F_3 plants that need evaluation for actual HIR through resource-intensive haploid induction and haploid selection. Chen et al. (2020) also reported the effectiveness of simultaneous MAS for *qhir1* and *qhir8*, resulting in a substantial increase in HIR by 3–14% and the elimination of approximately 90% of low-HIR genotypes.

Our study did not include inducer families with *qhir8* only because we aimed to investigate the synergistic effects between *qhir1* and *qhir8* on HIR. Previous studies have reported that *qhir8* alone resulted in poor or even null HIR. For instance, Chen et al. (2020) reported that the HIR of the plants with *qhir8* only ranged from 0.70% to 1.04%, which was significantly lower than either those that carried a heterozygous *qhir1* allele or those that carried a homozygous *qhir1*, with HIRs of 3.77% to 5.27% and 10.02% to 14.42%, respectively.

Previous studies reported six minor QTL (*qhir2, qhir3, qhir4, qhir5, qhir6, and qhir7*) (Prigge et al., 2012) and a novel gene, *ZmPLD3* (Li et al., 2021). Mutations of the *ZmPLD3* gene resulted in a haploid induction rate (HIR) comparable to that of the homozygous recessive *MTL* gene. This mutation showed synergistic effects rather than functional redundancy in tripling HIR in the presence of the homozygous recessive *MTL* gene. Later

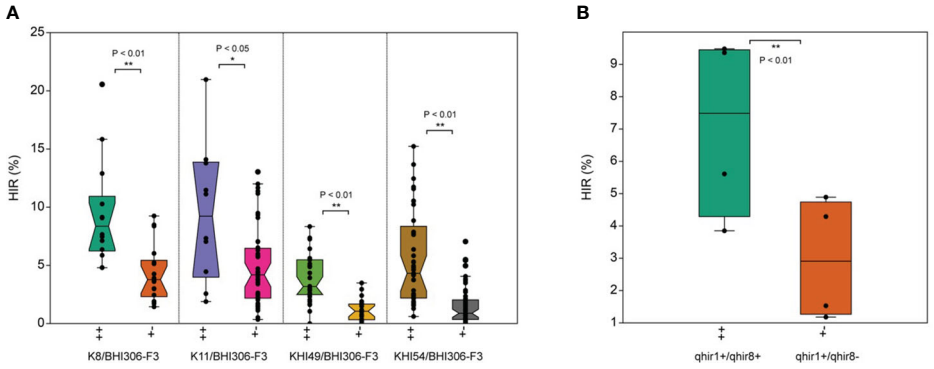


FIGURE 2
Haploid induction rate (HIR) in F_3 inducer families with *qhir1+/qhir8+* (++) and *qhir1+/qhir8–* (+–) genotypes involving **(A)** four different inducer parents; and **(B)** averaged means over four respective inducer populations. *, and ** data significant through paired samples t-test at $P<0.05$ and $P<0.01$, respectively.

in 2022, Meng and colleagues manipulated the Stock6-derived inducer lines by overexpressing maize *CENH3* fused with different fluorescent protein tags and found that the engineered Stock6-derived lines showed a noticeable increase in the maternal HIR up to 16.3%, which was increased by ~6.1% than Stock6-derived lines control (Meng et al., 2022). Hu et al. (2016) found two minor QTL responsible for HIR expression, namely *qhir11* and *qhir12*, which are closely linked to the major QTL *qhir1*. While the *qhir11* was not diagnostic for differentiating inducers and non-inducers, the *qhir12* had a haplotype allele common to all inducer lines observed but not found in all non-inducers studied. In addition, they noticed that the *qhir12* region was related to three candidate genes involved in DNA or amino acid binding. (Nair et al., 2017) performed a genome wide association study (GWAS) and identified more than 20 SNPs associated with HIR in two different association mapping panels. A recent genome-wide association study (GWAS) involving 159 haploid inducers has confirmed the polygenic nature of HIR and identified a major gene near *MTL*, a significant QTL on chromosome 10, and other minor QTL on six of the ten chromosomes (Trentin et al., 2023a). It is conceivable that these QTL, or even undiscovered ones, may be present in our inducer genotypes, highlighting the need for further investigations to discover novel QTL conferring HIR. Drawing insights from Prigge et al. (2012), this endeavor is feasible, as the number of QTL and the magnitude of QTL effects for HIR can vary across populations and generations.

TABLE 2 The mean comparison between *qhir1+/qhir8+* and *qhir1+/qhir8–* genotypes of each F_3 family on haploid induction rate (HIR, %).

Population name	Gene combination	HIR (%)	PNU	TPN	*p*-value	PVE (%)
K8/BHI306-F_3	*qhir1+/qhir8+*	9.36 A	14	32	1.05E−03 **	36
	qhir1+/qhir8–	4.29 a	18			
K11/BHI306-F_3	*qhir1+/qhir8+*	9.48 A	10	54	4.01E−02 *	17
	qhir1+/qhir8–	4.89 a	44			
KHI-49/BHI306-F_3	*qhir1+/qhir8+*	3.85 B	31	52	6.21E−08 **	39
	qhir1+/qhir8–	1.18 b	21			
KHI-54/BHI306-F_3	*qhir1+/qhir8+*	5.61 B	41	99	1.91E−07 **	33
	qhir1+/qhir8–	1.53 b	58			

PNU the number of plants; TPN the total number of plants; PVE proportion of variance in phenotypes explained (%).
HIR means (%) followed by different uppercase letters within the same *qhir1+/qhir8+* genotype are significantly different based on Tukey's Honestly Significant Difference (HSD) Test at 5%.
HIR means (%) followed by different lowercase letters within the same *qhir1+/qhir8–* genotype are significantly different based on Tukey's Honestly Significant Difference (HSD) Test at 5%.
* and ** data significant through paired samples t-test at $P<0.05$ and $P<0.01$, respectively.

This present study serves as a continuation of the haploid inducer breeding program, focusing on achieving high HIR and local adaptation to the tropical savanna in Thailand. In our previous reports, relying solely on phenotypic selection in the breeding strategy did not yield promising haploid inducers with satisfactory HIR, i.e., below 6% in two families KHI49 and KHI54 (Dermail et al., 2021) and two populations K8 and K11 (Thawarorit et al., 2023). The incorporation of genetic recombination with BHI306, an elite inducer stock carrying favorable alleles for HIR, and the implementation of precise selections such as MAS for simultaneous loci have now enabled us to obtain promising inducer genotypes. Notably, some individual plants within *qhir1+/qhir8+* genotypes in families K8/BHI306 and K11/BHI306 demonstrated HIRs exceeding 20%, surpassing both founder parents (Supplementary Table S1).

The significant variations for HIR among families within the same qhir1+/qhir8+ genotype (Table 2) imply the importance of the genetic background of founder parents to establish those inducer families. We noticed that families KHI-49/BHI306 and KHI-54/BHI306 had significantly lower abilities to induce haploids than families K8/BHI306 and K11/BHI306. Although the four females (KHI-49, KHI-54, K8, and K11) derived from the same haploid inducer, Stock-6, they experienced different selection schemes. Regarding the pedigree information, the females KHI-49 and

10.3389/fpls.2024.1337463

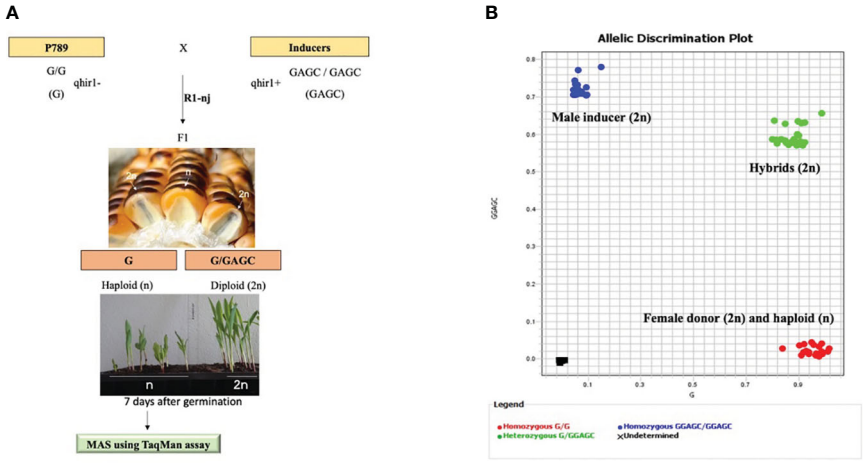


FIGURE 3
Workflow for haploid identification through the *R1-nj* and haploid validation by MAS using *qhir1* TaqMan assay **(A)** and the allelic distribution plot from TaqMan assay **(B)**. The diploid and haploid classes are represented by the green and red dots, respectively. Genotype BHI306 was used as male inducer, whereas genotype P789 was used as female donor. The SNP graphs for other populations can be found in Supplementary Figure S2.

KHI-54 had experienced long-term selections, including six for non-HIR related traits and the following three for *R1-nj* kernel marker. Some favorable alleles responsible for HIR may be lost during selections since Chaikam et al. (2019) argued that non-inducer pollen showed selection advantages over inducer pollen. In contrast, the females K8 and K11 only experienced one selection cycle among F_2 populations derived from crosses between Stock-6 haploid inducer and non-inducer waxy maize germplasm. We assumed that the proportion of HIR-related favorable alleles in the K8 and K11 genotypes was higher than in KHI-49 and KHI-54.

TABLE 3 Ploidy identification (haploid vs. diploid) via TaqMan assay for *qhir1* in a sub-set population of induction crosses between a male inducer BHI306 and a female donor P789.

No	Population name	*qhir1*+ (GGAGC/GGAGC)	*qhir1* +/*qhir1*– (GGAGC/G)	*qhir1* – (G/G)	Total
1	BHI306 – male inducer	10	0	0	10
2	KHI49/BHI306-F_3 – male inducer	31	0	0	31
3	KHI54/BHI306-F_3 – male inducer	41	0	0	41
4	P789 – female donor	0	0	7	7
5	P789/BHI306-F_1 – putative haploid	0	1	10	11
6	P789/BHI306-F_1 – putative diploid	0	10	0	10
7	P789/(KHI49/BHI306-F_3)-F_1 – putative haploid	0	10	146	156
8	P789/(KHI49/BHI306-F_3)-F_1 – putative diploid	0	27	0	27
9	P789/(KHI54/BHI306-F_3)-F_1 – putative haploid	0	13	162	175
10	P789/(KHI54/BHI306-F_3)-F_1 – putative diploid	0	28	0	28
Total		82	89	325	496

Plant samples with *qhir1*+/*qhir1*– are defined as true diploids.
Plant samples with *qhir1*– are defined as true haploids.
Putative haploid and diploid are based on the preliminary haploid identification via the *R1-nj* marker.

Although *per se* on HIR can be altered by different testing environments and donor germplasm (Kebede et al., 2011; Prigge et al., 2011; De La Fuente et al., 2018; Sintanaparadee et al., 2022), our current finding suggests the presence of transgressive segregants in F_3 families. We recommend further phenotypic evaluations in inducer families with *qhir1*+/*qhir8*+ genotypes, focusing on traits related to the ideotype of haploid inducers, such as plant height, ear height, flowering behaviors, tassel and pollen attributes, and seed set. This assessment will help determine the feasibility of those genotypes in haploid induction stage, whether in induction nurseries or isolation fields (Trentin et al., 2020; Trentin et al., 2023b). Considering the polygenic nature governing HIR and those mentioned agronomic traits, genomic selection approach can be applied to simultaneously identify promising parents to generate progenies with favorable performance on targeted traits prior to field evaluation. Almeida et al. (2020) implemented genomic prediction for cross prediction and parental selection in a haploid inducer breeding program with varying levels of accuracy depending on traits evaluated and suggested that HIR and agronomic traits can be improved simultaneously.

In our study, we proved that MAS for *qhir1* is effective to confirm the true-to-type of haploids. The induced progenies were clustered into two pools according to haplotypes. This allelic clustering can be explained by two hypotheses: (1) single fertilization occurs when only the egg or the central cell is fertilized, resulting in kernels with haploid embryos (Sarkar and Coe, 1966) and (2) selective elimination of inducer genomes from embryonic cells (Zhao et al., 2013). Acknowledging the small sample sizes used, Linnet (1999) suggested that the minimum sample size for optimizing the regression analysis should fall

withing the range of 40 to 100 samples. Therefore, conducting further replicated trials with a larger sample size is encouraged before fully realizing the potential of this approach in haploid identification in maize. As a practical proposal, molecular markers could be employed to verify *R1-nj*-based putative haploids at the early seedling stage, not exceeding seven days after planting (DAP). This timeline aligns with the common practice of haploid genome doubling using colchicine at 10-12 DAP (Vanous et al., 2017). To prevent the risk of *R1-nj* marker misclassification, an additional phenotypic marker, the red root phenotype at seedling stage from *Pl-1* gene, was used. This phenotype results from light-independent anthocyanin production, although exposure to light conditions can induce anthocyanin pigmentation for some genotypes (Coe, 1994). Moreover, oil content was used as a screening criterion for haploid and diploid using nuclear magnetic resonance (NMR) (Wang et al., 2016). The success of this method required high-oil haploid inducers (Liu et al., 2022). Preventing high false positives through molecular assays can help to reduce the DH line production costs, because false positives can be discarded prior to haploid genome doubling (Baleroni et al., 2021).

Conclusions

Our study revealed that implementing marker-assisted selection (MAS) for *qhir1* and *qhir8* at an early generation (F_3) substantially enhanced the haploid induction rate (HIR) of tropical × temperate haploid inducer families. On average, the HIRs of families homozygous for both *qhir1*+ and *qhir8*+ were 1–3-fold higher than those homozygous for *qhir1*+ only. The *qhir1* marker, utilizing the TaqMan assay, effectively distinguished diploid/haploid progenies at the early seedling stage (≤7 DAP) with high accuracy (100%), as validated by flow cytometric analysis. We propose the integration of MAS to expedite the breeding of haploid inducers for high HIR, complementing the use of the *R1-nj* marker for the identification of true haploids.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Author contributions

KK: Writing – original draft, Conceptualization, Data curation, Formal analysis, Methodology. AD: Writing – original draft, Conceptualization, Methodology, Writing – review & editing. KS: Writing – review & editing, Conceptualization, Funding acquisition, Methodology, Supervision. TL: Supervision, Writing – review & editing. SW: Supervision, Writing – review & editing. BT: Supervision, Writing – review & editing. WP: Supervision, Writing – review & editing. TT: Supervision, Writing – review & editing. VR: Conceptualization, Methodology, Supervision, Writing – review & editing. SA: Writing – review & editing, Supervision.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The National Science and Technology Development Agency (NSTDA) (Grant No. P-20-52286, P-21-50610 and P-23-51489). Also, the National Science, Research and Innovation Fund, Thailand Science Research and innovation (TSRI).

Acknowledgments

The authors would like to thank the Plant Breeding Research Center for Sustainable Agriculture, Faculty of Agriculture, Khon Kaen University, Thailand, for providing plant materials and research facilities. As well as the High-Quality Research Graduate Development Cooperation Project between Kasetsart University and the National Science and Technology Development Agency (NSTDA) for providing the scholarship.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2024.1337463/full#supplementary-material

References

Almeida, V. C., Trentin, H. U., Frei, U. K., and Lübberstedt, T. (2020). Genomic prediction of maternal haploid induction rate in maize. *Plant Genome.* 13, e20014. doi: 10.1002/tpg2.20014

Baleroni, A. G., Ré, F., Pelozo, A., Kamphorst, S. H., Carneiro, J. W. P., Rossi, R. M., et al. (2021). Identification of haploids and diploids in maize using seedling traits and flow cytometry. *Crop Breed. Appl. Biotechnol.* 21, e38422145. doi: 10.1590/1984-70332021v21n4a54

Barret, P., Brinkmann, M., and Beckert, M. A. (2008). Major locus expressed in the male gametophyte with incomplete penetrance is responsible for *in situ* gynogenesis in maize. *Theor. Appl. Genet.* 117, 581–594. doi: 10.1007/s00122-008-0803-6

Chaikam, V., Martinez, L., Melchinger, A. E., Schipprack, W., and Boddupalli, P. M. (2016). Development and validation of red root marker-based haploid inducers in maize. *Crop Sci.* 56, 1678–1688. doi: 10.2135/cropsci2015.10.0653

Chaikam, V., Molenaar, W., Melchinger, A. E., and Boddupalli, P. M. (2019). Doubled haploid technology for line development in maize: technical advances and prospects. *Theor. Appl. Genet.* 132, 3227–3243. doi: 10.1007/s00122-019-03433-x

Chaikam, V., Nair, S. K., Babu, R., Martinez, L., Tejomurtula, J., and Boddupalli, P. M. (2015). Analysis of effectiveness of *R1-nj* anthocyanin marker for *in vivo* haploid identification in maize and molecular markers for predicting the inhibition of *R1-nj* expression. *Theor. Appl. Genet.* 128, 159–171. doi: 10.1007/s00122-014-2419-3

Chaikam, V., Nair, S. K., Martinez, L., Lopez, L. A., Utz, H. F., Melchinger, A. E., et al. (2018). Marker-assisted breeding of improved maternal haploid inducers in maize for the tropical/subtropical regions. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01527

Chen, C., Xiao, Z., Zhang, J., Li, W., Li, J., Liu, C., et al. (2020). Development of *in vivo* haploid inducer lines for screening haploid immature embryos in maize. *Plants.* 9, 739. doi: 10.3390/plants9060739

Coe, E. H. (1994). ""Anthocyanin genetics,"," in *The maize handbook* (Springer, New York, NY, USA), 279–281.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. *Euphytica* 142, 169–196. doi: 10.1007/s10681-005-1681-5

Cooper, M., Gho, C., Leafgren, R., et al. (2014). Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. *J. Exp. Bot.* 65, 6191–6204. doi: 10.1093/jxb/eru064

De La Fuente, G. N., Frei, U. K., Trampe, B., et al. (2018). A diallel analysis of a maize donor population response to in *vivo* maternal haploid induction: I. Inducibility. *Crop Sci.* 58, 1830–1837. doi: 10.2135/cropsci2017.05.0285

Dermail, A., Chankaew, S., Lertrat, K., Lübberstedt, T., and Suriharn, K. (2021). Selection gain of maize haploid inducers for the tropical savanna environments. *Plants.* 10, 2812. doi: 10.3390/plants10122812

Dermail, A., Lübberstedt, T., Suwarno, W. B., Chankaew, S., Lertrat, K., Ruanjaichon, V., et al. (2023). Combining ability of tropical× temperate maize inducers for haploid induction rate, *R1-nj* seed set, and agronomic traits. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1154905

Doležel, J. (1997). Application of flow cytometry for the study of plants genomes. *J. Appl. Genet.* 38, 285–302.

Dong, L., Li, L., Liu, C., Liu, C., Geng, S., Li, X., et al. (2018). Genome editing and double-fluorescence proteins enable robust maternal haploid induction and identification in maize. *Mol. Plant* 11, 1214–1217. doi: 10.1016/j.molp.2018.06.011

Dong, X., Xu, X., Li, L., Liu, C., Tian, X., Li, W., et al. (2014). Marker-assisted selection and evaluation of high oil *in vivo* haploid inducers in maize. *Mol. Breed* 34, 1147–1158. doi: 10.1007/s11032-014-0106-3

Geiger, H. H., and Gordillo, G. A. (2009). Doubled haploids in hybrid maize breeding. *Maydica* 54, 485.

Gilles, L. M., Khaled, A., Laffaire, J. B., Chaignon, S., Gendrot, G., Laplaige, J., et al. (2017). Loss of pollen-specific phospholipase NOT LIKE DAD triggers gynogenesis in maize. *EMBO J.* 36, 707–717. doi: 10.15252/embj.201796603

Hu, H., Schrag, T. A., Peis, R., Unterseer, S., Schipprack, W., Chen, S., et al. (2016). The genetic basis of haploid induction in maize identified with a novel genome-wide association method. *Genetics.* 202, 267–1276. doi: 10.1534/genetics.115.184234

Jacquier, N. M. A., Gilles, L. M., Martinant, J. P., Rogowsky, P. M., and Widiez, T. (2021). Maize in planta haploid inducer lines: A cornerstone for doubled haploid technology. *Doubled Haploid Technol.* 2, 25–48. doi: 10.1007/978-1-0716-1335-1_2

Kebede, A. Z., Dhillon, B. S., Schipprack, W., Araus, J. L., Bänziger, M., Semagn, K., et al. (2011). Effect of source germplasm and season on the *in vivo* haploid induction rate in tropical maize. *Euphytica.* 180, 219–226. doi: 10.1007/s10681-011-0376-3

Kelliher, T., Starr, D., Richbourg, L., Chintamanani, S., Delzer, B., Nuccio, M. L., et al. (2017). MATRILINEAL, a sperm-specific phospholipase, triggers maize haploid induction. *Nature.* 542, 105–109. doi: 10.1038/nature20827

Koebner, R. M. (2004). Marker assisted selection in the cereals: the dream and the reality. *Cereal Genomics*, 317–329. doi: 10.1007/1-4020-2359-6_10

Li, Y., Lin, Z., Yue, Y., et al. (2021). Loss-of-function alleles of *ZmPLD3* cause haploid induction in maize. *Nat. Plants.* 7, 1579–1588. doi: 10.1038/s41477-021-01037-2

Linnet, K. (1999). Necessary sample size for method comparison studies based on regression analysis. *Clin. Chem.* 45, 882–894. doi: 10.1093/clinchem/45.6.882

Liu, C., Li, J., Chen, M., Li, W., Zhong, Y., Dong, X., et al. (2022). Development of high-oil maize haploid inducer with a novel phenotyping strategy. *Crop J.* 10, 524–531. doi: 10.1016/j.cj.2021.07.009

Liu, C., Li, X., Meng, D., Zhong, Y., Chen, C., Dong, X., et al. (2017). A 4-bp insertion at ZmPLA1 encoding a putative phospholipase A generates haploid induction in maize. *Mol. Plant* 10, 520–522. doi: 10.1016/j.molp.2017.01.011

Meng, D., Luo, H., Dong, Z., Huang, W., Liu, F., Li, F., et al. (2022). Overexpression of modified CENH3 in Maize Stock6-derived inducer lines can effectively improve maternal haploid induction rates. *Front. Plant Sci.* 13, 892055. doi: 10.3389/fpls.2022.892055

Nair, S. K., Molenaar, W., Melchinger, A. E., Boddupalli, P. M., Martinez, L., Lopez, L. A., et al. (2017). Dissection of a major QTL qhir1 conferring maternal haploid induction ability in maize. *Theor. Appl. Genet.* 130, 1113–1122. doi: 10.1016/j.cj.2019.09.008

Nanda, D. K., and Chase, S. S. (1966). An embryo marker for detecting monoploids of maize (*Zea mays* L.). *Crop Sci.* 6, 213–215. doi: 10.2135/cropsci1966.0011183X000600020036x

Pfosser, M., Amon, A., Lelley, T., and Heberlebors, E. (1995). Evaluation of sensitivity of flow-cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. *Cytometry* 21, 387–393. doi: 10.1002/cyto.990210412

Prasanna, B. M. (2012). Diversity in global maize germplasm: Characterization and utilization. *J. Biosci.* 37, 843–855. doi: 10.1007/s12038-012-9227-1

Prigge, V., Sánchez, C., Dhillon, B. S., Schipprack, W., Araus, J. L., Bänziger, M., et al. (2011). Doubled haploids in tropical maize: I. Effects of inducers and source germplasm on *in vivo* haploid induction rates. *Crop Sci.* 51, 1498–1506. doi: 10.2135/cropsci2010.10.0568

Prigge, V., Xu, X., Li, L., Babu, R., Chen, S., Atlin, G. N., et al. (2012). New insights into the genetics of *in vivo* induction of maternal haploids, the backbone of doubled haploid technology in maize. *Genetics.* 190, 781–793. doi: 10.1534/genetics.111.133066

Qiu, F., Liang, Y., Li, Y., Liu, Y., Wang, L., and Zheng, Y. (2014). Morphological, cellular and molecular evidences of chromosome random elimination *in vivo* upon haploid induction in maize. *Curr. Plant Biol.* 1, 83–90. doi: 10.1016/j.cpb.2014.04.001

Sarkar, K. R., and Coe, E. H. (1966). A genetic analysis of the origin of maternal haploids in maize. *Genetics.* 54, 453–464. doi: 10.1093/genetics/54.2.453

Sintanaparadee, P., Dermail, A., Lübberstedt, T., Lertrat, K., Chankaew, S., Ruanjaichon, V., et al. (2022). Seasonal variation of tropical savanna altered agronomic adaptation of Stock-6-derived inducer lines. *Plants.* 11, 2902. doi: 10.3390/plants11212902

Thawarorit, A., Dermail, A., Lertrat, K., Chankaew, S., and Suriharn, K. (2023). Stratified haploid identification system through the *R1-nj* kernel and reduced seedling vigor in tropical maize germplasm. *Biodiversitas.* 24, 4262–4268. doi: 10.13057/biodiv/d240807

Trentin, H. U., Batîru, G., Frei, U. K., Dutta, S., and Lübberstedt, T. (2022). Investigating the effect of the interaction of maize inducer and donor backgrounds on haploid induction rates. *Plants.* 11, 1527. doi: 10.3390/plants11121527

Trentin, H. U., Frei, U. K., and Lübberstedt, T. (2020). Breeding maize maternal haploid inducers. *Plants.* 9, 614. doi: 10.3390/plants9050614

Trentin, H. U., Krause, M. D., Zunjare, R. U., Almeida, V. C., Peterlini, E., Rotarenco, V., et al. (2023a). Genetic basis of maize maternal haploid induction beyond *MATRILINEAL* and *ZmDMP*. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1218042

Trentin, H. U., Yavuz, R., Dermail, A., Frei, U. K., Dutta., S., and Lübberstedt, T. (2023b). A comparison between inbred and hybrid maize haploid inducers. *Plants.* 12, 1095. doi: 10.3390/plants12051095

Vanous, K., Vanous, A., Frei, U. K., and Lübberstedt, T. (2017). Generation of maize (*Zea mays*) doubled haploids via traditional methods. *Curr. Protoc. Plant Biol.* 2, 147–157. doi: 10.1002/cppb.20050

Wang, H., Liu, J., Xu, X., Huang, Q., Chen, S., Yang, P., et al. (2016). Fully-automated high-throughput NMR system for screening of haploid kernels of maize (corn) by measurement of oil content. *PloS One* 11, e0159444. doi: 10.1371/journal.pone.0159444

Xu, X., Li, L., Dong, X., Jin, W., Melchinger, A. E., and Chen, S. (2013). Gametophytic and zygotic selection leads to segregation distortion through *in vivo* induction of a maternal haploid in maize. *J. Exp. Bot.* 64, 1083–1096. doi: 10.1093/jxb/ers393

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: From publications to practice. *Crop Sci.* 48, 391–407. doi: 10.2135/cropsci2007.04.0191

Xu, Y., McCouch, S. R., and Zhang, Q. (2005). How can we use genomics to improve cereals with rice as a reference genome? *Plant Mol. Biol.* 59, 7–26. doi: 10.1007/s11103-004-4681-2

Zhao, X., Xu, X., Xie, H., Chen, S., and Jin, W. (2013). Fertilization and uniparental chromosome elimination during crosses with maize haploid inducers. *Plant Physiol.* 163, 721–731. doi: 10.1104/pp.113.223982

Zhong, Y., Liu, C., Qi, X., Jiao, Y., Wang, D., Wang, Y., et al. (2019). Mutation of *ZmDMP* enhances haploid induction in maize. *Nat. Plants.* 5, 575–580. doi: 10.1038/s41477-019-0443-7

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 18 April 2024
DOI 10.3389/fpls.2024.1319938

OPEN ACCESS

EDITED BY
Baohua Wang,
Nantong University, China

REVIEWED BY
Stephen Welch,
Kansas State University, United States
Shawn Carlisle Kefauver,
University of Barcelona, Spain

*CORRESPONDENCE
Eva M. Molin
eva-maria.molin@ait.ac.at

†These authors have contributed equally to this work and share first authorship

RECEIVED 11 October 2023
ACCEPTED 13 March 2024
PUBLISHED 18 April 2024

CITATION
Chang-Brahim I, Koppensteiner LJ, Beltrame L, Bodner G, Saranti A, Salzinger J, Fanta-Jende P, Sulzbachner C, Bruckmüller F, Trognitz F, Samad-Zamini M, Zechner E, Holzinger A and Molin EM (2024) Reviewing the essential roles of remote phenotyping, GWAS and explainable AI in practical marker-assisted selection for drought-tolerant winter wheat breeding. *Front. Plant Sci.* 15:1319938. doi: 10.3389/fpls.2024.1319938

Reviewing the essential roles of remote phenotyping, GWAS and explainable AI in practical marker-assisted selection for drought-tolerant winter wheat breeding

Ignacio Chang-Brahim[1†], Lukas J. Koppensteiner[2†], Lorenzo Beltrame[3†], Gernot Bodner[4], Anna Saranti[5], Jules Salzinger[3], Phillipp Fanta-Jende[3], Christoph Sulzbachner[3], Felix Bruckmüller[3], Friederike Trognitz[1], Mina Samad-Zamini[2], Elisabeth Zechner[6], Andreas Holzinger[5] and Eva M. Molin[1,5*]

[1]Unit Bioresources, Center for Health & Bioresources, AIT Austrian Institute of Technology, Tulln, Austria, [2]Saatzucht Edelhof GmbH, Zwettl, Austria, [3]Unit Assistive and Autonomous Systems, Center for Vision, Automation & Control, AIT Austrian Institute of Technology, Vienna, Austria, [4]Department of Crop Sciences, Institute of Agronomy, University of Natural Resources and Life Sciences Vienna, Tulln, Austria, [5]Human-Centered AI Lab, Department of Forest- and Soil Sciences, Institute of Forest Engineering, University of Natural Resources and Life Sciences Vienna, Vienna, Austria, [6]Verein zur Förderung einer nachhaltigen und regionalen Pflanzenzüchtung, Zwettl, Austria

Marker-assisted selection (MAS) plays a crucial role in crop breeding improving the speed and precision of conventional breeding programmes by quickly and reliably identifying and selecting plants with desired traits. However, the efficacy of MAS depends on several prerequisites, with precise phenotyping being a key aspect of any plant breeding programme. Recent advancements in high-throughput remote phenotyping, facilitated by unmanned aerial vehicles coupled to machine learning, offer a non-destructive and efficient alternative to traditional, time-consuming, and labour-intensive methods. Furthermore, MAS relies on knowledge of marker-trait associations, commonly obtained through genome-wide association studies (GWAS), to understand complex traits such as drought tolerance, including yield components and phenology. However, GWAS has limitations that artificial intelligence (AI) has been shown to partially overcome. Additionally, AI and its explainable variants, which ensure transparency and interpretability, are increasingly being used as recognised problem-solving tools throughout the breeding process. Given these rapid technological advancements, this review provides an overview of state-of-the-art methods and processes underlying each MAS, from phenotyping, genotyping and association analyses to the integration of explainable AI along the entire workflow. In this context, we specifically address the challenges and importance of breeding winter wheat for greater drought tolerance with stable yields, as

regional droughts during critical developmental stages pose a threat to winter wheat production. Finally, we explore the transition from scientific progress to practical implementation and discuss ways to bridge the gap between cutting-edge developments and breeders, expediting MAS-based winter wheat breeding for drought tolerance.

KEYWORDS

drought tolerance, GWAS, MAS, plant breeding, winter wheat, XAI, UAV remote phenotyping, smart agriculture

1 Introduction

Water scarcity is seen as a key threat for the 21st century (Unesco, 2012), with global water demand expected to surpass supply by 40% by 2030 (Gilbert, 2010). Even under the 'Green Path' Shared Socioeconomic Pathway (SSP1), which envisions a future with increased sustainability and reduced resource and energy consumption (Riahi et al., 2017), Europe is projected to experience a rise in the maximum annual temperature of over 5°C, a decrease in precipitation of about -700 mm, and a reduction in soil water content of up to -62 kg/m^2 by 2060 relative to 2020 (see Figure 1A). Given that agriculture is the primary user of freshwater, accounting for 70% of total withdrawal globally (FAO, 2010; Hoekstra and Mekonnen, 2012), it is crucial to develop new strategies to enhance crop water use efficiency through agronomy or breeding to tackle the impending water crisis (Sposito, 2013; Turner et al., 2014; Bodner et al., 2015).

The yield of wheat, one of the key staple crops worldwide and particularly in Western Europe (about 14% and 25% of total cropland area respectively; FAO, 2023), has seen a steady increase during the second half of the 20th century. However, this trend has shifted since the 1990s, with yields reaching a peak and partially even slightly decreasing, and showing an increasing variability year-to-year. Brisson et al. (2010) suggested two main factors for this shift: (i) the effects of climate change and (ii) a decrease in input intensity, primarily of N-fertiliser, due to EU agri-environmental regulations. Therefore, future production of key crops like wheat will have to cope with higher resource constraints, in terms of both water and nutrients, even in Europe's temperate climate conditions. Particularly in sub-humid to semi-arid regions, the balance between soil water supply and crop water demand largely determines achievable yield levels (Lalic et al., 2013). With projected higher temperatures and more unpredictable rainfalls, the frequency of periods of crop water shortage is likely to increase (Qin et al., 2023). Additionally, the co-occurrence of heat and drought is expected to have the most significant impact on wheat yield, with a predicted global reduction of 3.9% (Heino et al., 2023). On a more regional scale, for instance, climate change projections for the Pannonian lowlands, an important wheat-producing region in Europe, indicate that the number of dry days with water deficit during the vegetation period will increase (Trnka et al., 2011; Lalic et al., 2013; Schils et al., 2018; van der Velde et al., 2018). Evaluation of past yield data and simulation model predictions point to a high risk for wheat production under climatic conditions with hot temperatures (>25°C; Lüttger and Feike, 2018; Figure 1B) and drought occurring at a sensitive developmental stage, such as germination, tillering, flowering or grain filling (Yu et al., 2018; Senapati et al., 2021; Xu et al., 2022). These factors underscore the urgency to speed up the breeding process for more drought (and heat) tolerant varieties to keep pace with the rate and scale of climate change.

One of the methods that has revolutionised plant breeding by improving its efficiency, speed and precision is marker-assisted selection (MAS) (Collard and Mackill, 2008). There are different MAS strategies such as MAS backcrossing, MAS pyramiding or early generation MAS (Jeon et al., 2023), all of which use DNA-based markers to help select lines with the desired traits. The limited number of markers per trait and its restricted use for traits under complex genetic control are major limitations of MAS. These limitations led to the development of other marker-based strategies such as genomic selection (GS) or crop growth models (Budhlakoti et al., 2022; Zhang et al., 2022). Unlike MAS, GS uses all available (genome-wide) markers to calculate a breeding value and has been shown to outperform MAS in several studies (Arruda et al., 2016; Degen and Müller, 2023). Despite these advancements, MAS is still extensively used to efficiently screen for traits of interest. For instance, MAS has been employed in wheat breeding to improve resistance to biotic and abiotic stresses and to maintain yield potential (Song et al., 2023; Subedi et al., 2023). A notable advantage of MAS may be that, compared to the genome-wide approach of GS, only a few markers ultimately need to be used by the breeder, making MAS –despite its limitations– an affordable solution for practical breeding.

However, for a MAS programme to be successful, certain prerequisites must be met: the generation of high-quality phenotypic and genotypic data, the understanding of marker-trait associations, the characterisation of reliable markers and, finally, the development of cost-efficient and easy-to-use genotyping approaches. In this review, we therefore attempt to cover this process using the example of winter wheat breeding for increased drought tolerance. As a starting point, (i) we revisit the physiological mechanisms and corresponding traits that have been associated with drought tolerance in winter wheat under different drought regimes (Section 2), (ii) we further discuss traditional and modern phenotyping approaches focusing on airborne

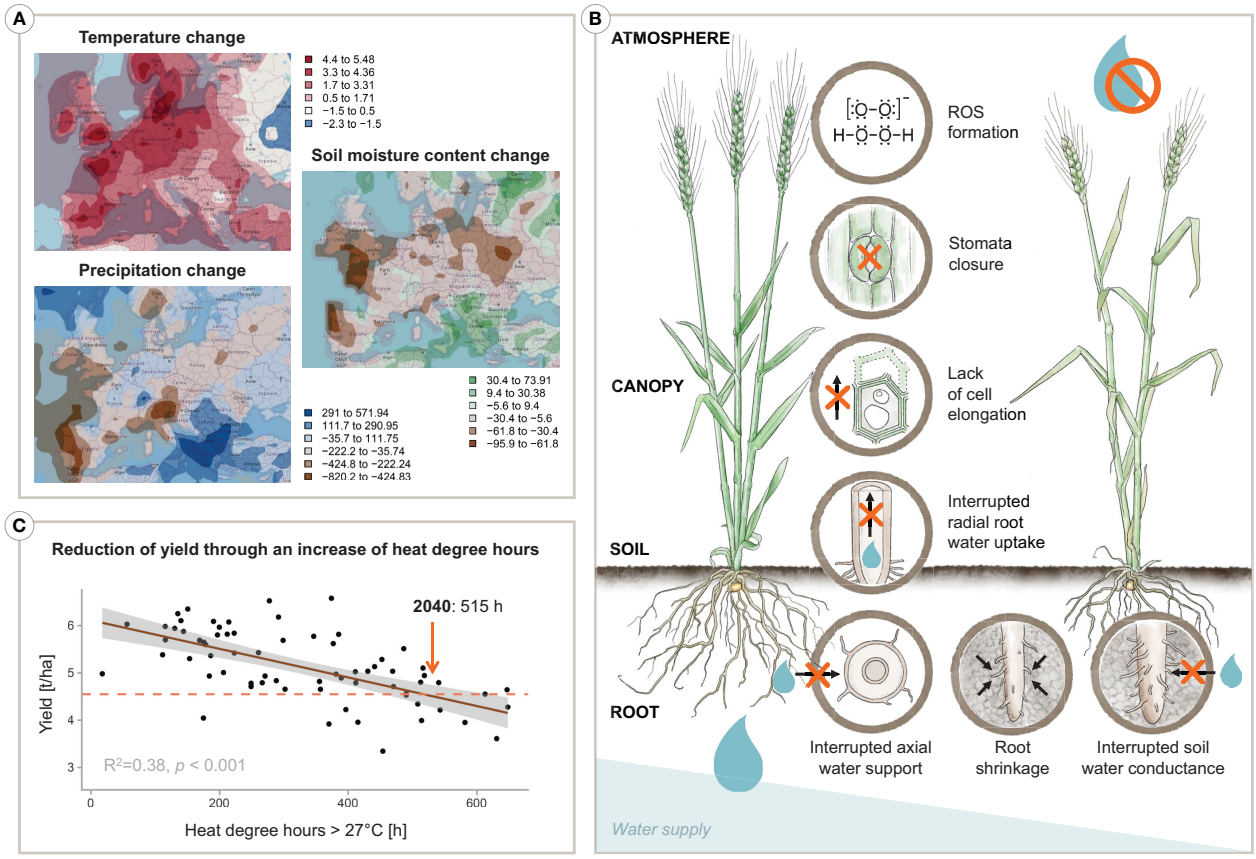


FIGURE 1
Infographic depicting **(A)** climatic projections by 2060 in Europe based on the SSP1 scenario, **(B)** an increase of heat degree hours results in a decrease of yield in wheat, and **(C)** a plant's various physiological responses to water deprivation. Specifically, the projections in **(A)** show the change of the annual maximum temperature [°C], the change of precipitation [mm], and the change of the soil moisture content [kg/m^2] by 2060 relative to 2020 considering the best case SSP1 narrative following the 'Green Path'. In detail, maps are based on the GFDL-ESM4 model data provided by NOAA-GFDL, release year 2018 (Krasting et al., 2018) representing the SSP1-2.6 model made available through the Coupled Model Intercomparison Project, CMIP (Eyring et al., 2016). Furthermore, in **(B)**, a decrease of wheat yield [tonnes/ha] can be seen with rising heat degree hours over the vegetation period. On-farm yield data and heat degree hours represent averages of six districts in Lower Austria during the years 2002–2014. The dashed line indicates a projection to 2040. An overview of the physiological reactions of a plant to drought stress is presented in **(C)**. Designed by Tatjana Hirschmugl and Eva M. Molin.

technologies and time series records and provide a guide for airborne data acquisition for winter wheat (Section 3), (iii) we include genome-wide association studies (GWAS), an important computational approach that links the recorded phenotypes with the genotypes for the identification of genetic markers used in MAS (Sections 4 and 5), and finally, (iv) we address artificial intelligence (AI) models accompanied by explainable AI (xAI) methods that could support the breeding process at several steps in the context of smart agriculture (Section 6). Attempting to bridge the gap between scientific innovations and their application in practice, (v) we conclude this review with an overview of the practical work of plant breeders (Section 7) and where these (novel) cutting-edge approaches could fit in and help accelerate the breeding process.

2 Physiological mechanisms underlying drought tolerance

Historically, advances in wheat breeding have largely been driven by increased yield potential through better assimilate partitioning towards grain sinks, sustained by prolonged assimilate source activity due to extended green canopy duration (Lichthardt et al., 2020). However, under water-limited conditions, yield formation is a complex function of total water uptake, water use efficiency, and harvest index (Passioura, 1977). Ecophysiological theory has guided trait-based breeding by uncovering stress adaptation strategies in natural vegetation. Levitt's scheme of dehydration avoidance, dehydration tolerance, and drought escape (Levitt, 1980) serves as a guiding framework for physiological breeding: plant traits underlying individual stress response types aid targeted selection for crop adaptation in water-limited environments (e.g., Richards, 2006; Araus et al., 2008; Cattivelli et al., 2008).

The selection of relevant traits involved in drought tolerance mechanisms that could potentially lead to better and more stable yields strongly depends on the time when the drought occurs (van Ginkel et al., 1998; Blum, 2011). For instance, the phenological adaptation ('drought escape') of early maturity might be especially sensitive to early drought events while thriving in summer-dry regions with water deficiency during the grain-filling stage. Dehydration avoidance by 'water saving' (Levitt, 1980) might

result in suboptimal use of available water under moderate drought regimes, while in situations with more severe drought and crop growth largely dependent on stored soil moisture from off-season rainfalls, a 'conservative' water use preserves water for grain filling and yield formation (Mori et al., 2011).

As highlighted in Figure 1C, the regulation of plant water balance forms the physiological basis for identifying potential breeding traits for more drought-tolerant plants. Whether transpiration can meet the potential demand, driven by the atmospheric vapour pressure deficit, depends not only on the availability of soil water but also on the transport capacity of soil and plants under variable driving gradients (e.g., Maseda and Fernandez, 2006). In coarse to medium-textured soils, the transport of water through the tortuous soil pore system to the root surface drops sharply when larger pores drain upon successive soil drying, resulting in supply limitation (wilting) at a water content substantially higher than the permanent wilting point (Czyż et al., 2012). With successive drying, the root-soil contact can be lost due to root shrinkage and air gap formation as well as root mucilage becoming hydrophobic to protect root tissues from dehydration (Carminati and Vetterlein, 2013; Affortit et al., 2023). Stomata are the ultimate regulators of crop water transport, providing a mechanism to prevent plants from dehydration damage (cf. Figure 1C). Stomata thus act upon imbalances between vapour losses from and liquid water transport to the transpiring leaves. Root water uptake (Abdalla et al., 2022) and xylem transport (Cruiziat et al., 2002) are crucial for stomata regulation, mediated by chemical and hydraulic signals within plant-specific safety margins (Sperry and Love, 2015). Sustained xylem water flow under high-pressure gradients between soil and atmosphere without interruption of transport vessels by air embolism, leading to an eventual hydraulic failure of the transport system, has been suggested as one of the key bottlenecks for crop performance in dry environments (Sperry et al., 1998; Vadez et al., 2013; Vadez, 2014). Plants relying on high safety margins with sensitive stomata response to tissue dehydration (isohydric behaviour; Tardieu and Simonneau, 1998; Hochberg et al., 2018), also have to cope with increased leaf temperature and high radiation load at the leaf, which leads to an overproduction of reactive oxygen species that cause metabolic disorders and limit plant growth and development (Mukarram et al., 2021). Within this general framework of physiological mechanisms and related traits, Blum (2009) points to maximising water uptake as a focus for breeding because it is generally compatible with high yields, i.e. genotypes that fall into the category of 'water wasters' according to Levitt's framework. Efficient water uptake by the root system is a desirable breeding objective (Vadez et al., 2007). In wheat, physiological and root research studies indicate a significant contribution of the root system to increased drought tolerance (e.g. MansChadi et al., 2008; Palta et al., 2011; Becker et al., 2016; Li et al., 2021).

To expand the germplasm sources of (novel) stress tolerance traits, landraces and crop wild relatives are a valuable resource offering a wealth of diversity (Galluzzi et al., 2020) that could be transferred into breeding programmes, as has been extensively reviewed for wheat (Valkoun, 2001; Reynolds et al., 2006; Trethowan and Mujeeb-Kazi, 2008; Nakhforoosh et al., 2015; Lehnert et al., 2022; Aloisi et al., 2023; Shokat et al., 2023). Specifically, cereal genetic resources could contribute to improved drought tolerance through higher water use efficiency (Konvalina et al., 2010), rapid early development (Mullan and Reynolds, 2010), stem reserve demobilisation, osmotic adjustment (Reynolds et al., 2006), and even plant waxiness (Patidar et al., 2023). Several studies also suggest a contribution of root traits (e.g., Reynolds et al., 2006; Sanguineti et al., 2007; Trethowan and Mujeeb-Kazi, 2008; Lopes and Reynolds, 2010; Nakhforoosh et al., 2014).

Despite these studies, further progress in physiological and trait-based breeding to accelerate wheat improvement for future environmental conditions critically depend on adequate selection strategies that combine (advanced) targeted trait phenotyping (see Section 3) with modern genetic tools (see Sections 4 and 5).

3 From traditional to airborne phenotyping

The practice of measuring phenotypic traits dates back to Neolithic agriculture when domesticated cereals were intentionally selected for traits such as broad kernels (Zohary et al., 2012). Today, one of the cornerstones of plant breeding is the selection of superior individuals based on phenotypic traits (e.g., grain yield), and more recently, the identification of genome regions controlling these traits (cf. Sections 4 and 5). With advancements in sensor technology, phenotyping has evolved into a high-throughput process, including remote sensing and machine learning (ML), offering solutions for precision agriculture and digital plant breeding (Walter et al., 2015; Pieruschka and Schurr, 2019; Holzinger et al., 2022a; Jeon et al., 2023). This diversity of phenotyping approaches is mirrored in the wide range of data and data formats obtained during the breeding process by different sensors (Thoday-Kennedy et al., 2022), such as visual scorings, direct measurements of plant phenotypic parameters, meteorological readings, and hyperspectral and multispectral measurements (Heremans et al., 2015; Adão et al., 2017; Becker and Schmidhalter, 2017; Hu et al., 2020; Saranya et al., 2023), which we aim to cover in this review with respect to wheat.

3.1 Traditional phenotyping

Modern plant breeding still depends on traditional phenotyping, which includes visual scorings, plant measurements, and destructive sampling followed by laboratory analysis (Furbank and Tester, 2011; Atkinson et al., 2018). Each type has unique characteristics in terms of precision and measurement speed. Non-destructive measurements are easily measured, such as plant height and visual assessments of disease occurrence, phenology, and plant architecture. These visual assessments are commonly used and are also applied for official national variety testing, e.g., in Austria (Kumar et al., 2016; Steiner et al., 2017; Anderegg et al., 2020; AGES, 2023; Lunzer et al., 2023). However, the precision of non-destructive measurements can be limited by various factors such as observer variability and lighting conditions. Conversely, destructive measurements involve the

collection and analysis of plant samples to acquire data on above-ground dry matter, grain yield, and quality traits like protein content and baking quality. Despite offering high precision, these measurements are time-consuming, destructive, and often limited by cost considerations.

As for breeding experimental setups, they can be classified based on the degree of control over environmental conditions (Hammer and Hopper, 1997). Growth chambers provide highly controlled conditions, where numerous environmental variables such as temperature, light intensity, and CO_2 concentration can be manipulated (Rezaei et al., 2018). Semi-controlled conditions, observed in, e.g., greenhouses and rain-out shelters, offer some control over environmental factors, with greenhouses affording greater control than rain-out shelters (Yadav, 2017; Rezaei et al., 2018). Finally, experiments under field conditions feature the lowest control over environmental variables. Nevertheless, field experiments are undoubtedly relevant, since most of them are conducted in the field under uncontrolled conditions (Hammer and Hopper, 1997). They allow for scientific testing of experimental factors under conditions similar to agriculture practice. Experimental factors can include varying genotypes, sowing times, fertilisation, plant protection, irrigation and disease occurrence due to natural pressure as well as artificial inoculation (Buerstmayr et al., 2000; Koppensteiner et al., 2022).

Observational units vary across setups, ranging from plots in field experiments and rain-out shelters to pots in greenhouses and growth chambers (Buerstmayr et al., 2000; Yadav, 2017; Rezaei et al., 2018; Koppensteiner et al., 2022). In field trials, units of observation include single seeds (Zhu et al., 2012), single rows (Buerstmayr et al., 2000), micro-plots (Miedaner et al., 2006), and large plots (Koppensteiner et al., 2022), e.g., 1.5 m by 7 m, depending on the amount of available seed material of a genotype in the respective stages of the breeding process. In the context of UAV-based sensor systems discussed in this review, micro-plots and large plots are the most relevant observation units. Measurements on more detailed levels are possible depending on the specifications of sensor systems and operational flight height. Despite the significance of field experiments, conducting field phenotyping is arduous, time-intensive, and susceptible to human and environmental variability. Therefore, there is a pressing need to enhance field phenotyping capabilities to facilitate accurate and high-throughput phenotyping, thus expediting crop breeding processes (Yang et al., 2020).

3.2 Remote sensing

Remote phenotyping techniques in digital agriculture are prized for their non-destructive nature and their ability to improve data collection accuracy and efficiency (Sishodia et al., 2020; Jeon et al., 2023). These techniques rely on remote sensing, which involves detecting electromagnetic radiation across various wavelengths emitted, reflected, or transmitted by objects. Remote sensing measurements are categorised into direct and indirect methods. Direct measurements involve directly gauging traits of interest, such as plant height using digital surface and terrain models (Holman et al., 2016), while indirect measurements estimate traits using statistical or ML models like biomass and water stress estimates (Wang et al., 2016; Das et al., 2021).

Remote phenotyping can be conducted at various scales: ground-based - handheld or vehicle-mounted (Kumar et al., 2020; Tang et al., 2023), aerial - via aircraft or UAVs (Fei et al., 2023; Nguyen et al., 2023), and satellite platforms like Sentinel-2 (Zhao et al., 2020; European Space Agency, 2023a), Landsat (Zhou et al., 2020; Darra et al., 2023; NASA, 2023), WorldView-2 and 3 (Tattaris et al., 2016; Yuan et al., 2017; European Space Agency, 2023b), or RapidEye (Eitel et al., 2007; European Space Agency, 2024). To contextualise these platforms, key remote sensing features are spatial, temporal, spectral, and radiometric resolutions (Verde et al., 2018). Spatial resolution refers to pixel size, temporal resolution to the time between measurements, spectral resolution to the number of spectral channels, and radiometric resolution to a sensor's ability to detect varying energy quantities in a specific spectral channel. Each phenotyping platform presents trade-offs; for instance, ground-based techniques offer high spatial resolution but require dedicated manpower, leading to lower time resolution. Aerial technologies offer enhanced operational performance and sub-centimetre spatial resolution (Bhandari et al., 2020) but are weather-dependent, limiting time-series data availability. Satellites provide densely populated time series but sacrifice spatial resolution, with modern satellites offering resolutions as low as 31 cm in the case of Worldview-3 (European Space Agency, 2023b). Moreover, in general, increasing sensor-object distance or increasing the swath width of the satellite, i.e. the horizontal distance covered by a satellite sensor, can improve temporal resolution by allowing the sensor to revisit the same location more frequently. However, this enhancement comes at the cost of diminished spatial resolution (Kadhim et al., 2016). Other trade-offs do not depend on the spatial resolution, but, for UAV, the maximum weight of a payload determines the equipped camera and therefore the spectral resolution available to be measured (Mohsan et al., 2023).

Another key concept in remote sensing and therefore in remote phenotyping is the Ground Sampling Distance (GSD), i.e. the spatial spacing between the centres of two consecutive pixels as measured on the ground. It is determined by several key factors: altitude (h), denoting the height above the ground at which the sensor is positioned and affecting the scale of the captured image; sensor size (s), representing the physical size of the sensor in the camera, typically measured in mm, larger sensors capturing more detail and impacting the GSD; focal length (f), the distance from the optical centre of the lens to the camera sensor, measured in mm, influencing the field of view and magnification of the captured image, and image resolution (r). The GSD is mathematically represented as:

$$\text{GSD (m)} = \frac{h\,\text{(m)} \times s\,\text{(mm)}}{f\,\text{(mm)} \times r\,\text{(pixels)}}$$

This metric is important because it directly determines the spatial resolution of the imagery, affecting the level of detail that can be captured and the accuracy of any measurements or analyses

conducted on the images. Generally, a smaller GSD indicates a higher spatial resolution and finer detail in the imagery. GSD values vary across different imaging platforms. For UAV imaging, GSD can vary depending on factors like altitude and sensor specifications, generally falling between 0.5 to 10 cm per pixel (Yuan et al., 2018). This range allows for moderately detailed aerial imagery suitable for various agricultural and environmental applications. On the other hand, satellite imaging offers broader coverage but typically lower spatial resolution. GSD for satellite imagery can range from 30 cm to several m per pixel, depending on the satellite platform, sensor, and imaging mode employed (Chawade et al., 2019).

In plant breeding, the field experimental plot is the typical unit of observation (Hammer and Hopper, 1997). While current satellite systems' spatial resolution may be inadequate for precise phenotypic parameters at a plot level (Tattaris et al., 2016), ground-based and aerial remote sensing approaches offer suitable spatial resolution. UAVs, with their flexibility, extended operational times, lower cost, and high spatial resolution in the low centimetre range, emerge as promising phenotyping platforms for plant breeding and precision agriculture (Sishodia et al., 2020; Guo et al., 2021).

3.3 UAV-based remote phenotyping

UAV remote sensing coupled with ML provides a non-destructive method that enables repeated plant measurements over time. This is a significant improvement over traditional methods, which are laborious, time-consuming and expensive (Galieni et al., 2021; Nguyen et al., 2023). Therefore, the use of UAVs for remote phenotyping has become a well-established practise in plant breeding (Yang et al., 2020; Guo et al., 2021). Compared to other remote sensing platforms, UAVs offer several advantages. They are capable of swiftly collecting spectral data, outperforming the speed of handheld devices. They can capture data at a higher resolution compared to aerial cameras operated from a manned aircraft, and they are not dependent on satellite overpasses for data collection in the region of interest (Kim et al., 2019).

3.3.1 An overview of UAV sensor systems

UAVs can be equipped with passive sensors, such as multispectral, hyperspectral and thermal cameras, or active sensors, such as Light Detection And Ranging (LiDAR) (Thoday-Kennedy et al., 2022).

Since multispectral and hyperspectral cameras can capture data at various wavelengths (also outside the visible spectrum), their use in agricultural applications offers many benefits. They can identify and monitor crop health and stress (Yang et al., 2009; Virnodkar et al., 2020), determine and map corn emergence uniformity (Vong et al., 2022) and quickly detect diseases and pests (Prabhakar et al., 2012). Multispectral leaf reflectance data are very useful because they contribute to computing indices widely used in agriculture (see Table 1 for an overview of the main vegetation indices). Additionally, both multispectral and hyperspectral data can be utilised to estimate crop yields using ML methods (Fei et al., 2023; Joshi et al., 2023). In contrast to multispectral sensors, which typically capture broader spectral bands with spectral resolutions from 10 to 100 nm, hyperspectral sensors offer a much higher spectral resolution, often within 1 to 10 nm (Adão et al., 2017). They effectively capture a spectral continuum across hundreds of contiguous, narrow bands, enabling detailed pixel-by-pixel analysis. Hyperspectral cameras are capable of capturing not only the visible (400-700 nm) and near-infrared (NIR, 700-2500 nm) wavelength ranges but also radiation from the ultraviolet (UV, 100-400 nm) to thermal infrared (TIR, 3000-15000 nm) wavelengths. However, the large data storage required for hyperspectral data can limit its use in large-scale applications (Sun et al., 2019). Therefore, despite their significant advantages, hyperspectral applications in large-scale wheat phenotyping could face challenges related to data storage, management, and budget constraints (Ang and Seng, 2021).

Thermal imaging, which operates within the broader long-wave infrared (LWIR) wavelengths (from 8 to 15 μm), serves as a valuable tool for detecting plant stress. Thermal measurements can be used to evaluate the transpiration status, plant vigour, and the spread of diseases in wheat cultivars (Mahlein et al., 2012) or, together with measurement of the air temperature, to compute the Crop Water Stress Index (CWSI). This index can then be incorporated as a feature in an ML model to provide insights on canopy head evapotranspiration or to segment image data into temperature areas (Zhou et al., 2021b). Moreover, combining thermal imaging data with other phenotypic traits improves the holistic understanding of plant responses to environmental conditions. This synergy enables researchers, breeders and farmers to make well-informed decisions for optimal crop management and resource allocation (Khanal et al., 2017; Stutsel et al., 2021).

On the other hand, unlike camera-based systems that passively capture reflected, transmitted, or emitted light, LiDAR is an active technique that emits laser pulses and measures the time for these pulses to reflect off objects, providing precise distance and spatial data. This has been particularly useful in wheat breeding for estimating plant biomass and plant height (Hütt et al., 2023). Taking advantage of global navigation satellite systems (GNSS) and laser altimetry, and using GIS software, accurate crop height measurements can be obtained by subtracting a digital terrain model from a digital surface model representing the crop canopy surface (Jenal et al., 2021). Although LiDAR systems typically operate at a single wavelength, combining geometric measurement with spectral information is possible, such as registering multispectral camera images with LiDAR point clouds (Hakula et al., 2023), or using LiDAR systems with individual lasers at various frequencies, e.g., Optech Titan (GEO3D, 2023).

3.3.2 Spectral indices supporting smart wheat breeding

In the context of wheat breeding, an index is a mathematical formula designed to provide a comprehensive representation of

TABLE 1 Most used indices of remote phenotyping applied in wheat breeding.

Name	Formula	Properties	Reference
NDVI Normalised Difference Vegetation Index	$\frac{NIR - R}{NIR + R}$	Vegetation density, plant health, and land cover monitoring	(Carlson and Ripley, 1997)
EVI Enhanced Vegetation Index	$G_F \frac{NIR - R}{NIR + C_1R - C_2B + L}$	Sensitivity in high vegetation areas and aerosol correction	(Matsushita et al., 2007)
SAVI Soil Adjusted Vegetation Index	$\frac{(NIR - R)}{(NIR + R + L)}(1 + L)$	Vegetation index corrected for Soil Condition	(Huete, 1988)
NDWI Normalised Difference Water Index	$\frac{NIR - SWIR}{NIR + SWIR}$	Water presence detection and water content sensitivity	(Gao, 1996)
LAI Leaf Area Index	$\frac{-\ln P(\theta)\cos(\theta)}{G(\theta)\Omega(\theta)}$	Green leaf area measurement and ecosystem dynamics monitoring	(Nilson, 1971)
TCARI Transformed Chlorophyll Absorption in Reflectance Index	$3 \cdot (R_{700} - R_{670}) - 0.2(R_{700} - R_{550})\frac{R_{700}}{R_{670}}$	Chlorophyll estimation in vegetation	(Haboudane et al., 2002)
GNDVI Green Normalised Difference Vegetation Index	$\frac{NIR - G}{NIR + G}$	Vegetation monitoring	(Rahman and Robson, 2016)
MSAVI Modified Soil Adjusted Vegetation Index	$\frac{2NIR + 1 - \sqrt{(2NIR + 1)^2 - 8(NIR - R)}}{2}$	Enhanced sensitivity to low vegetation	(Qi et al., 1994)
ARI Anthocyanin Reflectance Index	$ARI = R_{550}^{-1} - R_{700}^{-1}$	Detection of plant pigments	(Gitelson et al., 2001)
NDRE Normalised Difference Red Edge	$\frac{NIR - Red\ \ edge}{NIR + Red\ \ edge}$	Measurement of vegetation stress	(Tilling et al., 2007)
CCCI Canopy Chlorophyll Content Index	$\frac{NDRE - NDRE_{min}}{NDRE_{max} - NDRE_{min}}$	Measurement of chlorophyll content in the canopies	(Fitzgerald et al., 2010)

The LAI formula presented here is not the only one available. Other methods for computing LAI are referenced in Fang et al. (2019). SAVI/EVI: G_F is a gain factor, C_1 and C_2 are the coefficients to correct for aerosol influences in the red band and L is the Canopy background adjustment factor. LAI: $P(\theta)$ represents the canopy gap fraction at the zenith angle θ and $G(\theta)$ is the projection function corresponding to the fraction of foliage projected on the plane normal to the solar direction and $\Omega(\theta)$ is the canopy clumping index. ARI/TCARI: The term RY typically denotes the measurement of the red colour at a wavelength denoted by Y, with the unit of measurement being nanometers (nm). CCCI: $NDRE_{min}$ and $NDRE_{max}$ represent the minimum and maximum values of NDRE that have been recorded, respectively.

various plant traits, physiological states and characteristics (Reynolds and Langridge, 2016). It combines different desired traits into a single numerical value, allowing breeders to assess and compare the overall performance of different wheat varieties more thoroughly (Myneni et al., 1995). The computation of these indices creates a multidimensional profile, enriching the complexity of the breeding problem and providing valuable input for machine-learning approaches. Consequently, indices are crucial tools that enable breeders to make informed decisions, optimise their breeding strategies, and ultimately develop wheat varieties that thrive in a wide range of agricultural and environmental conditions in modern research (Radočaj et al., 2023).

In precision agriculture, vegetation indices are broadly categorised into two main types: broadband and narrowband (Thenkabail et al., 2002). Broadband indices, such as the Normalised Difference Vegetation Index (NDVI) (Rouse et al., 1974), integrate information from relatively wide spectral bands, such as the NIR band. These indices offer a generalised measure of vegetation vigour and health. This approach is efficient and simple, making these indices suitable for large-scale agricultural monitoring and management tasks where rapid assessment is prioritised. In contrast, narrowband indices, such as the Chlorophyll Absorption Ratio Index (TCARI) (Haboudane et al., 2002), target specific narrow spectral bands within the electromagnetic spectrum. These indices focus on precise absorption features related to chlorophyll content, leaf structure, and other biochemical properties. Narrowband indices provide high spectral resolution making them valuable for tasks requiring in-depth analysis of plant health and stress. The choice of using either family of indices depends on the specific physiological traits under investigation.

Table 1 presents several indices common in remote sensing for wheat phenotyping. The practical rationale behind our selection of these indices is the ease of computing them with standard commercially available multispectral cameras (NIR - 700-2500 nm, RGB - 400-700 nm, SWIR - 2500-3000 nm, Red Edge-700-730 nm) and their recognised impact in assessing the plant water status, general stress condition and phenological traits. Vegetation Indices play a crucial role in assessing various aspects of vegetation health and physiological traits. The Normalised Difference Vegetation Index (NDVI) is widely utilised due to its computational simplicity,

facilitating assessments of vegetation density, plant health, and water stress (Condorelli et al., 2018; Hassan et al., 2019; Huang et al., 2021). However, limitations such as computational approximations and instrument inaccuracies can occasionally hinder its effectiveness in evaluating plant stress (Khan et al., 2018).

To address these limitations, several alternative indices have been developed. The Enhanced Vegetation Index (EVI) enhances the vegetation signal in high biomass areas and corrects for aerosol factors (Khan et al., 2018). Additionally, the Soil-Adjusted Vegetation Index (SAVI) and Modified Soil-Adjusted Vegetation Index (MSAVI) correct for soil irradiation in areas with low canopy cover (Prudnikova et al., 2019). The Normalised Difference Water Index (NDWI) detects water presence and sensitivity to water content (Wu et al., 2009). The Green Normalised Difference Vegetation Index (GNDVI) specifically targets green vegetation, utilising the green band instead of red. Furthermore, the Normalised Difference Red Edge (NDRE) emphasises the red edge region of the spectrum instead of the red band. These last two indices correlate with leaf nitrogen content and are used for controlling nitrogen leaf status (Li et al., 2019).

For other physiological traits, specialised indices have been developed. The Transformed Chlorophyll Absorption Reflectance Index (TCARI) estimates chlorophyll content in vegetation and biomass (Wang et al., 2022). The Leaf Area Index (LAI) measures foliage density within a canopy by comparing leaf surface area to ground area. The Anthocyanin Reflectance Index (ARI) identifies the presence of anthocyanins, aiding in the assessment of plant stress, phenology, and disease infection (Koc et al., 2022). Lastly, the Canopy Chlorophyll Content Index (CCCI) estimates chlorophyll levels in vegetation by combining red and red edge bands (Cummings et al., 2021).

Each of these indices offers unique insights that can inform breeding decisions, including assessments of yield potential and drought resistance, thus necessitating careful selection among the myriad indices developed by the remote sensing community (Xue and Su, 2017).

3.3.3 Machine learning for interpreting high-throughput field phenotypic data

In these scenarios, ML techniques showcase their advantage over conventional approaches in predicting phenotypes (Ansarifar et al., 2021). As high-throughput phenotyping methods produce a large volume of data, the use of ML becomes pivotal in accurately interpreting and effectively leveraging this data, leading to more precise phenotype predictions (Shaikh et al., 2022). For example, Wang et al. (2016) presents how random forest (RF) models outperform simple multilayer perceptrons (MLPs) and support vector machines (SVMs) in predicting wheat biomass. Grinberg et al. (2020) provides a comparative study of different ML models on various phenotyping problems across different crops, including wheat. The advent of deep learning enhances the classification of crop images, offering unprecedented granularity in monitoring crop quality, assessing yield, and pinpointing water stress at a pixel-wise level (Chandel et al., 2021). Convolutional neural networks (CNNs) further boost the model's capabilities, automatically extracting key features and patterns to make reliable phenotype predictions (Jiang and Li, 2020). Moreover, deep learning models have expanded the range of possible predictions to include disease detection, stress severity quantification, and yield (Mohanty et al., 2016; Giménez-Gallego et al., 2019; Zhou et al., 2021a). An intriguing direction that research has taken is semi-supervised approaches to the learning problem (Tang et al., 2023; Zhou et al., 2023). Semi-supervised deep learning is an ML paradigm where a model is trained using a combination of labelled and unlabelled data. It uses the limited labelled data to guide the learning process and improve the model's performance on tasks such as classification or regression, while also benefiting from the larger pool of unlabelled data for generalisation and enhanced feature representation (Yang et al., 2021). Deep learning significantly improves the model's ability to generalise and enables accurate and reliable phenotyping models for high-throughput approaches. However, a key drawback of deep learning approaches is that each solution needs to be tailored to the data and the phenotypic trait under investigation.

While traditional methods continue to hold their merits, integrating (UAV-based) remote sensing coupled with ML in phenotyping processes might be essential to obtain better and more resilient crop varieties (Yang et al., 2020). In addition, operational costs could be significantly reduced by cutting fixed costs such as laboratory equipment and workforce. This would lead to improved scalability in the approach and quicker results that are passed over in the data pipeline.

3.4 A guide for UAV-based data acquisition for winter wheat

Moving to the next stage, this Section presents a detailed overview of a potential high-throughput field phenotyping system specifically tailored for winter wheat. The main objective is to facilitate the acquisition of phenotypic data for GWAS (see Section 4) and MAS techniques in the frame of precision agriculture. A schematic representation of the key components of the pipeline is presented in Figure 2.

In this scenario, the fundamental premise revolves around the division of the test field into georeferenced experimental plots, overseen by experts tasked with gathering *in-situ* data. This experimental arrangement mirrors established methodologies seen across various research endeavours, aimed at facilitating controlled crop cultivation (Bai et al., 2016; Haghighattalab et al., 2016; Volpato et al., 2021). The grid structure delineating individual plots within the field is visually represented in Figure 2. Typically, experts conduct assessments and record measurements by visually inspecting these plots, as demonstrated in Koc et al. (2022). It's also advantageous to conduct these measurements at specific intervals, tailored to the trait being studied. For instance, in Fernandez-Gallego et al. (2020), multispectral images were captured under direct sunlight on three dates: June 6th, 25th, and July 3rd, 2018, in Belgium, corresponding to the developmental stages flowering and ripening, to monitor wheat ear development and count. During each scheduled flight mission, a UAV systematically follows a

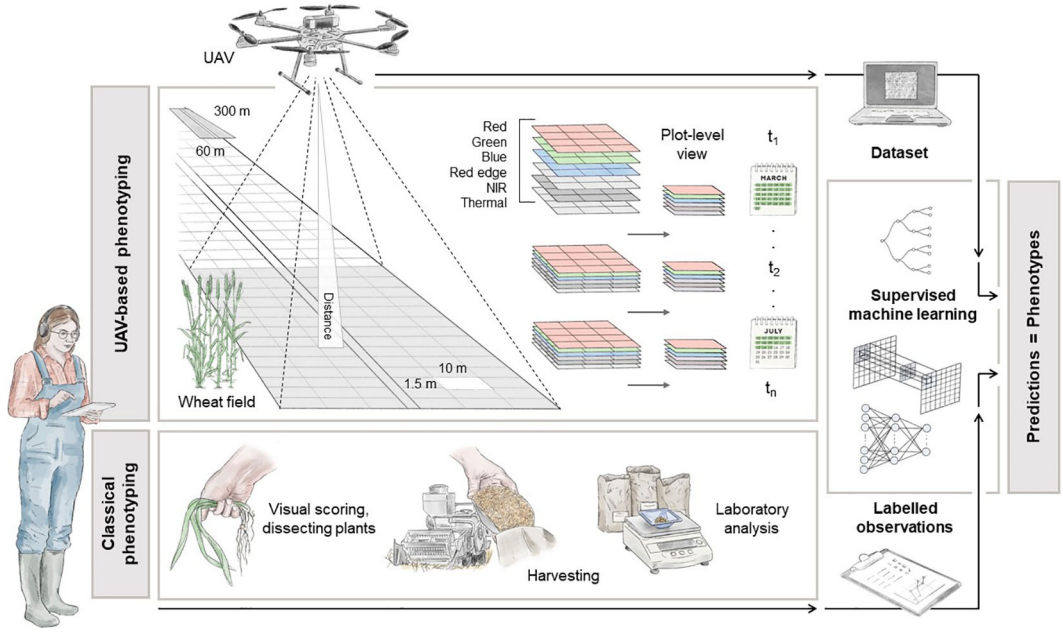


FIGURE 2
Illustration of a high-throughput UAV-based phenotyping configuration for plot-level analysis. The pipeline explains how to get from the raw data, *in-situ* phenotype acquisitions and raw images, to a structured and cohesive dataset to be employed in ML operations. The depicted dimensions, captured wavelengths, ML models, and traditional phenotyping methods serve as reference points. Designed by Tatjana Hirschmugl.

predefined grid pattern, meticulously gathering data while traversing the agricultural field.

The UAV could be equipped with a camera capable of capturing a range of spectral information, including RGB, panchromatic, Red Edge, NIR and thermal measures, during its flight (Holman et al., 2016; Tattaris et al., 2016; Duan et al., 2017). The specific selection of spectral bands depends on the particular index to be computed, which in turn depends on the trait under investigation. Additionally, the camera must undergo radiometric calibration to ensure the acquisition of physically meaningful measurements. The spatial resolution of the data acquired is influenced by both the altitude of the UAV and the intrinsic parameters of the camera used. For example, a standard multispectral camera (e.g. AgEagle Aerial Systems Inc, 2023) with 3.2 MP captures images with 2.5 cm GSD at an altitude of 60 m above the ground. The collected data is typically processed using photogrammetric software like Pix4D or Agisoft (Zhu et al., 2019). These software applications are used to create a reflectance map of the agricultural field by orthorectification and stitching individual images to reconstruct a high-resolution representation of the target area. Subsequently, plot-level spectral information is extracted using geospatial software, e.g. GIS, and organised for easy access (Beltrame et al., 2024). This data is then linked with specific plots and expert-acquired labelled information (lower part of Figure 2) to create tuples for subsequent ML analysis. These calibrated, cleaned, and standardised datasets can be used in classical preprocessing operations, including image normalisation, data augmentation, and sub-/oversampling techniques. To fully harness the information-rich content obtained, it is essential to select models that can handle the spatial complexity inherent in high-resolution images. For instance, a basic deep learning architecture, such as CNNs, can be used to extract feature maps from images and make accurate phenotype predictions (Kattenborn et al., 2021; Nguyen et al., 2023).

Recent advancements in image analysis, data extraction, and augmentation (Shorten and Khoshgoftaar, 2019), coupled with innovative artificial image synthesis techniques (Lu et al., 2022), and transfer learning (Hutchinson et al., 2017) are greatly enhancing the development and the integration of remote sensing technologies in agriculture. These advancements are starting to contribute to overcoming the phenotyping bottleneck (Song et al., 2021) and significantly enhance the provision of high-quality phenotype data to genotype - phenotype association studies ultimately resulting in an efficient and reliable MAS.

4 GWAS - a playground for the identification of genetic markers

Besides a meticulous recording of phenotypic data, MAS depends on the availability of genetic markers linked to the phenotypic trait of interest. Identifying these genetic regions associated with a phenotype is often not a straightforward task: many traits are polygenic, which adds to the complexity of their relationship with the phenotype (Korte and Farlow, 2013; Boyle et al., 2017; Mills and Rahal, 2019; Pierce et al., 2020). The general approach of linking genetic regions to traits, known as genetic mapping, consists of two main strategies: (biparental) linkage mapping (LM) and association mapping (AM) (March, 1999). LM utilises closely related individuals to study the co-segregation of markers and traits due to physical proximity, while AM uses

diverse, unrelated populations to detect statistical associations between markers and traits. AM, also known as linkage disequilibrium mapping, exploits linkage disequilibrium (Mackay and Powell, 2007), which is the nonrandom association pattern between alleles at different loci within a population (Nordborg and Tavaré, 2002; Gaut and Long, 2003). Since its introduction to plants (Tenaillon et al., 2001; Thornsberry et al., 2001), AM has become increasingly important in genetic research as cost-effective, high-throughput technologies for genotyping single nucleotide polymorphisms (SNPs) are now available, enabling dense marker coverage (Syvänen, 2005). A particular concept of AM, namely genome-wide association studies (GWAS), has become a common technique for understanding complex traits in plants in general and in many crop species, including wheat (Zhu et al., 2008; Cortes et al., 2021).

The primary advantage of GWAS is that it tests thousands to millions of genetic variants (e.g., SNPs) of many individuals from different populations on a genome-wide scale, allowing more complex genotype-phenotype relationships to be explained than with LM. However, for a genome-wide analysis, the knowledge about and the characterisation of SNPs is an essential part and is driven by the sequencing of the whole genome of the target organism. In the case of wheat, its genome was fully sequenced in 2018 (Appels et al., 2018) and has been continuously improved since then Shi and Ling (2018); Guan et al. (2020); Gao et al. (2023), including the creation of a pangenome (Montenegro et al., 2017; Jayakodi et al., 2021), which provides a valuable knowledge base for the development of a variety of high-density SNP arrays for high-throughput genotyping (Wang et al., 2014; Rimbert et al., 2018; Sun et al., 2020). Finally, to link these genotypic traits to the measured phenotypes, a wide range of GWAS-based tools and statistical methods are available, which have already been used in wheat, as shown in Table 2, which are described in the following Section in more detail.

TABLE 2 Common GWAS tools and methods, and examples of their application in wheat.

GWAS tool	Tool reference	GWAS method	Application in wheat
BayesCπ	Habier et al. (2011)	Bayesian GWAS	Zhao et al. (2013)
BLINK	Huang et al. (2019)	Bayesian GWAS	Devate et al. (2022)
EMMAX	Kang et al. (2010)	Efficient mixed model	Daba et al. (2018) Li et al. (2022)
farmCPU	Liu et al. (2016)	Multiple loci LMM	Gahlaut et al. (2021) Rahimi et al. (2023)
GAPIT	Lipka et al. (2012) and Tang et al. (2016)	Compressed mixed linear model based genomic prediction	Qaseem et al. (2019) Bennani et al. (2022)
MA	Zhou and Stephens (2012)	Genome-wide efficient mixed model	Wu et al. (2021) Ma et al. (2022)
JMP Genomics	SAS Institute Inc (2013)	GLMs	Gizaw et al. (2018)
PLINK	Purcell et al. (2007) and Chang et al. (2015)	Mixed model GWAS	Gogna et al. (2023) Zhao et al. (2023)
SNPtest	Marchini et al. (2007) and Marchini (2010)	Imputation based GWAS	Manickavelu et al. (2017) Muhammad (2021)
sommer	Covarrubias-Pazaran (2016)	Mixed model GWAS	Vukasovic et al. (2022) Dallinger et al. (2023)
TASSEL	Bradbury et al. (2007)	Generalised models and mixed linear models	Lehnert et al. (2018) Akram et al. (2021)

4.1 GWAS modelling strategies

The modelling strategies underlying GWAS are diverse from a statistical perspective, of which linear and Bayesian models are the prevailing strategies. Linear models fit linear equations to the data (genetic and phenotypic data), testing each specific marker and its relationship with the phenotype independently, simplifying the computational complexity that could arise from the genetic intricacies in the data (Sabatti, 2013). Generalised linear models (GLMs), as described in Nelder and Wedderburn (1972), add an additional layer of complexity, including a link function to relate input and output, thus providing certain flexibility from the rigidity of linearity. Linear mixed models (LMMs) represent another logical extension of linear models for GWAS and are widely applied (cf. Table 2). LMMs include fixed and random effects to model phenotypes, and can account for confounding factors such as population stratification, family structure, etc (Alamin et al., 2022). LMMs also offer versatility as they can analyse many experimental designs (Yang, 2010). These models, as their name suggests, assume a linear relationship between genotype and phenotype. They also assume that the random effects are normally distributed and that there is homoscedasticity in the variance of their errors (Warrington et al., 2014). These are the two main concepts use for GWAS methods based on linearity.

Bayesian models have also been developed and used for GWAS (cf. Table 2), fitting all markers simultaneously while addressing the issue of data dimensionality, making them well suited for polygenic traits (Fernando and Garrick, 2013; Miao et al., 2019). These methods require the specification of prior distributions, allowing knowledge of the data to be incorporated into them to yield more accurate results, with the caveat that deviation from the specified distribution can impair performance and statistical power (Cortes et al., 2021). Bayesian GWAS aim to identify sections of the genome that explain more than a threshold of the variance (Fernando and Garrick, 2013; Cortes et al., 2021). The multiple methods developed assume different

distributions for the calculation of the priors, having different performance according to the deviation from their actual distribution. Markov Chain Monte Carlo algorithms have been used to infer model parameters using Gibbs-type processes, as in Habier et al. (2011). The posterior probabilities of association, the odds of a specific SNP being actually related to the trait, can be calculated from the Bayes factor (Stephens and Balding, 2009).

4.2 Understanding the limitations of GWAS

Despite all these advancements, GWAS still have significant limitations in their design and application (Korte and Farlow, 2013; Wray et al., 2013; Tam et al., 2019; Cortes et al., 2021): they can be limited to the populations that are more represented in the studies, and there can exist a lack of transferability, as results may not extrapolate to other groups (Bouaziz et al., 2011), or the number of ostensible causal variations might be reduced if data from genetically diverse populations were used, so it is paramount to have an adequate representation of the population to reduce the possible biases that can arise from this (Clyde, 2019; Uffelmann et al., 2021). In addition, at this point, the causality or functionality of the linked SNPs is still elusive and only can be validated empirically through further experimentation (Hazelett et al., 2016; Gallagher and Chen-Plotkin, 2018). Non-normality of the data can also be a significant factor that increases error and reduces statistical power (Yoosefzadeh-Najafabadi et al., 2022). When applying GWAS, the risk of finding spurious correlations is ever-present, thus careful consideration must be taken into the model to correct when working with complex traits (Ball, 2013).

As the complexity of genetic architecture increases (Boyle et al., 2017), GWAS methods often fail to identify all genetic polymorphisms that have an effect on the phenotype. This phenomenon, known as missing heritability (Brachi et al., 2011), occurs when the genotype identified with these statistical methods does not fully explain the target characteristics. Missing heritability is thought to be caused partially by polymorphisms that have a small correlation with the target trait, and thus not being significant after Bonferroni correction (López-Cortegano and Caballero, 2019). Bonferroni correction is a method of adjusting p values when conducting multiple simultaneous tests on the same dataset; it involves dividing the initial p value by the number of hypotheses tested. In the context of GWAS, the relationship between specific SNPs and the desired trait is considered a comparison, so the p value is divided by the number of SNPs in the data (Napierala, 2012; Tam et al., 2019). However, Bonferroni correction has its drawbacks, for instance, when dealing with skewed phenotypic data (John et al., 2022). Since many GWAS methods are based on linear regression models, missing heritability could also be addressed with non-linear models (Peng, 2020). Nonetheless, some missing heritability might still be due to an underestimation of the effect sizes of common alleles, unidentified common and rare alleles, epigenetic changes, or in some cases, it might not even be found within genetic information (Marian, 2012; Bourrat et al., 2017). Colinearity is another potential source of reduced efficiency and statistical power for GWAS methodologies, and new strategies are needed to mitigate this limitation (Zhang et al., 2019). Finally, another important limitation of GWAS is high dimensionality of the data ($n \ll p$), where the number of features (e.g. SNPs) is much larger than the number of cases (e.g. genotypes), a common issue with biological data (Ramstein et al., 2019). Several AI concepts have been applied to overcome these limitations and disadvantages of GWAS (Szymczak et al., 2009; Nicholls et al., 2020; Enoma et al., 2022), some of which already include certain explainability (e.g., Mieth et al. (2021), see also Section 6).

Many target traits of GWAS are highly quantitative and complex. Grain yield and drought stress tolerance, for instance, are affected by interactions between underlying component traits (Allard and Bradshaw, 1964; Hammer et al., 2006). In Section 2, for instance, a wealth of physiological mechanisms that influence drought stress tolerance are presented. These interactions, however, can be non-linear (Chang and Zhu, 2017), which is a relevant challenge in GWAS. In this context, Technow et al. (2015) proposes to incorporate a crop growth model (CGM) directly into genomic analysis. Crop growth models can simulate biological and physical processes in agricultural systems including plants, environment and management (Holzworth et al., 2014). Relevant CGMs in this context need to include genotype-specific parameters (Oliveira et al., 2021). As a result, these models can capture the effects of non-linear interactions between underlying component traits on target traits (Technow et al., 2015). Gu et al. (2014), for instance, applied QTL mapping and the crop growth model GECROS to investigate the effect of genetic variation in leaf photosynthetic rate on crop biomass in rice. Furthermore, CGMs can help in identifying ideotypes to improve target traits and suitability to specific weather and management conditions (Chang et al., 2019; Bogard et al., 2021). Collins et al. (2021), for example, investigated drought adaptation in Australian wheat using the crop growth model APSIM and suggests limited-transpiration rate at high evaporative demand as a promising trait for selection by breeders.

5 GWAS to dissect drought tolerance in wheat

Despite its limitations, GWAS has become a crucial method for discovering loci for traits of interest, as discussed in the previous section. Drought is one of the most important abiotic stressors affecting wheat yield (Heino et al., 2023), prompting scientists and breeders to identify loci associated with drought stress tolerance.

In addition to grain yield *sensu stricto*, numerous other drought stress-related traits have been studied in wheat, including plant height and root architecture, as well as phenological traits like days to heading, anthesis or maturity (Mwadzingeni et al., 2016; Molero et al., 2019; Khadka et al., 2020; Saini et al., 2022). A summary of selected characterised markers and their associated traits in the context of drought tolerance in wheat, including the GWAS method used, is given in Table 3 and will be further detailed in the subsequent sections.

TABLE 3 Selected markers related to drought tolerance in wheat found with GWAS.

Selected trait (s)	Drought during	N° of Markers found	Important markers	Method (tool)	Reference
Leaf chlorophyll content	seedling stage	28	IWB26948	LMM (GAPIT)	Maulana et al. (2020)
Days to wilting	seedling stage	104	WPT-2356	LMM GLM (TASSEL)	Ahmed et al. (2021)
Grain yield and biomass	whole season	73	wsnp_Ex_Rep_c67786_66472676	LMM (GAPIT)	Bennani et al. (2022)
Grain yield	whole season	94	IWA5483	GLM (JMP Genomics)	Gizaw et al. (2018)
Grain yield	whole season	192	IAAV619, wsnp_Ex_c11120_18022932	LMM (TASSEL)	Suliman et al. (2021)
Grain yield	whole season	61	M7661	LMM GLM (TASSEL)	Akram et al. (2021)
Grain yield	whole season	37	M9766, M9769	Compressed LMM (GAPIT)	Mathew et al. (2019)
Grain yield	whole season	45	S7A_112977027	FarmCPU (NA)	Bhatta et al. (2018)
Grain yield	whole season	136	wsnp_BM134363A_Ta_2_4	LMM (GAPIT)	Qaseem et al. (2018)
Stress tolerance index	whole season	9	AX-111169510	LMM PCA + K GWAS (GAPIT)	Zhao et al. (2023)
Days to maturity	whole season	37	M1433, M6472, M1576	Compressed LMM (GAPIT)	Mathew et al. (2019)

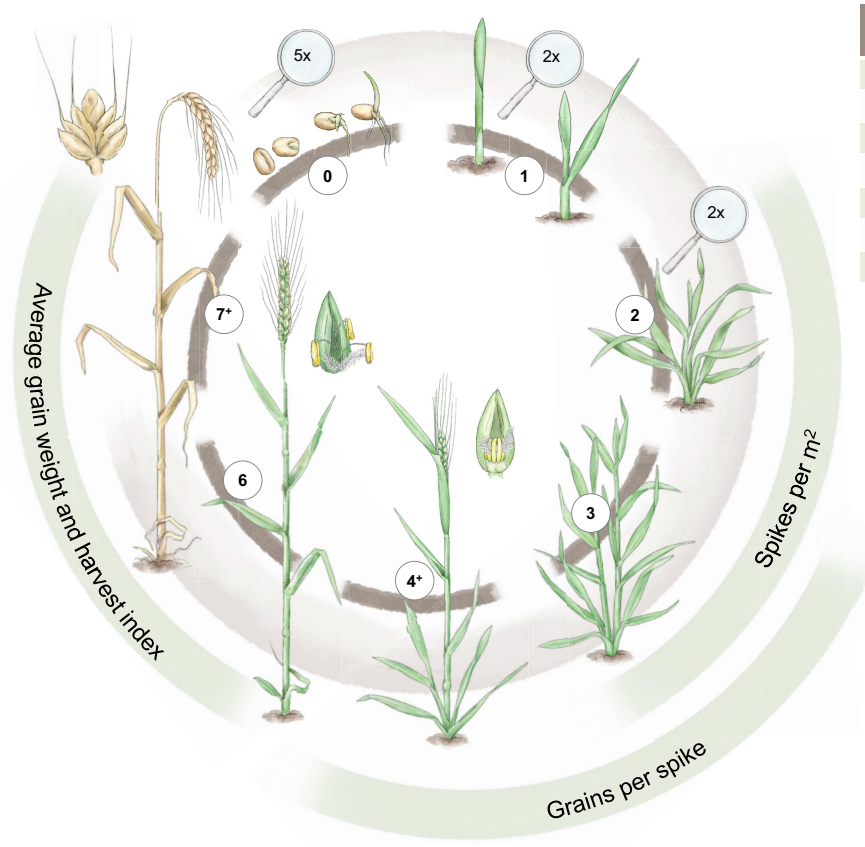
5.1 Introduction to developmental stages and yield components in wheat

To characterise marker-trait associations (MTAs) in the context of drought, it is essential to understand the developmental stages of wheat and to know at which stage drought can impact the traits of interest that might also affect grain yield, e.g. yield components, as highlighted in Figure 3. For the classification of the developmental stages, we use the commonly applied BBCH-code (Hack et al., 1992). Yield components are generally targets of high importance in plant breeding (Araus et al., 2008). In cereals, grain yield is described as the number of grains per m^2 multiplied by the average grain size. The number of grains per m^2 can be further differentiated into the number of spikes per m^2 and the number of grains per spike. Spikes per m^2 and grains per spike are established during the vegetative stage before anthesis, while the average grain size is mainly determined later during the generative stage (Geisler, 1983).

The number of spikes per m^2 is the first yield component determined during plant development. During the tiller differentiation process (BBCH 20, tillering stage, cf. Figure 3), the maximum number of tillers is established. Transitioning from BBCH 20 to BBCH 30 (stem elongation stage, cf. Figure 3), the number of established tillers is reduced to productive, spike-bearing tillers. Both the differentiation and reduction process of tillers are affected by drought stress. The tiller reduction process, however, is much more sensitive to water shortage than the respective differentiation process (Geisler, 1983). The differentiation process of generative organs, e.g., grains, can be divided into the establishment of spikelets and florets, whereby the primordia of spikelets are already developed by the end of tillering stage (BBCH 20). During stem elongation (cf. Figure 3), most spikelets and florets differentiate, and the maximum number of spikelets and florets is present at the beginning of BBCH 50 (heading). Afterwards, reduction processes of spikelets and florets occur until anthesis. The developmental stages from heading until anthesis are especially sensitive to drought stress González-Navarro et al. (2015). If drought stress is too severe, shedding of fertilised florets can occur after anthesis. Furthermore, insufficient water supply can also shorten the period for spikelet differentiation and thus reduce the number of spikelets per spike. In comparison to the simultaneous tiller reduction processes during stem elongation stage, this effect is minor (Geisler, 1983). Starting with anthesis (BBCH 60), the differentiation process of the caryopsis (the grain) occurs, which determines average grain weight (cf. Figure 3). In general, the longer the grain filling period during the stages of grain development and ripening (BBCH 70 and 80), the higher the average grain weight is (Ozturk et al., 2006; Klepeckas et al., 2020). The duration of this phase, however, is highly affected by environmental conditions. High temperature and insufficient water supply lead to shorter grain filling periods and thus a low average grain weight and even shrivelled grains (Spiertz, 1974; Klepeckas et al., 2020), as well as a shorter duration for the translocation of assimilates to the grain and thus lowers the harvest index of wheat (Davidson and Campbell, 2011; Neugschwandtner et al., 2015; Koppensteiner et al., 2022).

5.2 Markers associated to yield components under drought stress

As the processes of differentiation and reduction for each yield component occur at different developmental stages, they can be significantly affected by temporal environmental



Macro stage	Developmental stage	BBCH
0	Germination	0-9
1	Leaf development	10-19
2	Tillering	20-29
3	Stem elongation	30-39
4 & 5	Booting & heading	40-59
6	Anthesis	60-69
7, 8 & 9	Grain filling & ripening	70-99

FIGURE 3
The main developmental stages of winter wheat from germination to ripening are depicted. These developmental stages are contrasted by the most important yield components (spikes/m^2, grains/spike, and average grain weight and harvest index) as well as the BBCH classification in the adjacent table. The magnifying glasses indicate a magnification for the respective drawings. Designed by Tatjana Hirschmugl.

conditions (Satorre and Slafer, 2000). For instance, high temperature and water shortages can result in (i) accelerated plant development and consequently shorter differentiation processes for yield components, (ii) more intense reduction processes of individual yield components, and (iii) decreased photosynthetic activity, resulting in fewer available assimilates for grain filling. Besides environmental effects, yield components generally also depend on the genotype and crop management practices, such as sowing, fertilisation, plant protection, and irrigation (Geisler, 1983).

Numerous MTAs have already been characterised in experiments comparing wheat varieties and their responses to drought (Table 3). For example, Mathew et al. (2019) discovered associations between markers and biomass allocation to grain yield. Mwadzingeni et al. (2017) identified 334 MTAs with high confidence for traits under both drought and non-drought conditions. However, these markers explain only 20% of the phenotypic variation, which could be a consequence of the statistical stringency inherent in the methodology. The study found that chromosome 5 in genome D included QTLs related to grain yield, as seen in Quarrie et al. (2005). Among the 29 MTAs found for grain yield, some were located in genes annotated as F-box family protein or Sentrin-specific protease, described to have a potential role in drought stress tolerance (Bhatta et al., 2018). Markers such as *Xwmc273.3* and *Xpsp3094.1* have been used in the context of MAS of the yield-related QTL *Qyld.csdh.7AL* to develop high-yielding drought tolerant genotypes (Gautam et al., 2021). Bilgrami et al. (2020) identified SNPs (IWB39005 and IWB44377) related to the number of fertile tillers and total tillers. Suliman et al. (2021) explored grain yield and found 192 related markers, where 25 highly significant SNPs on chromosome 5A have a notable effect on grain yield, making this chromosome a relevant target for yield improvement under drought conditions. Seedling length, days to wilting, and leaf wilting were analysed in Ahmed et al. (2021), who reported 104 associated markers. Multiple phenotypic traits related to yield were used by Qaseem et al. (2018) for GWAS, resulting in 136 MTAs relevant for winter wheat's positive response to drought conditions.

5.3 Traits associated with phenology under drought conditions

It is well described that each developmental stage has its own specific water supply requirements. If drought occurs during water-sensitive developmental stages (cf. Figure 3), such as germination, tillering, flowering, or grain filling (Yu et al., 2018; Senapati et al., 2021; Xu et al., 2022), growth and subsequently yield can be

significantly impacted (Khadka et al., 2020). Therefore, the effects of the concurrence of critical phenological stages and drought conditions are critical (Langridge and Reynolds, 2021). Thus, in traditional plant breeding, phenological parameters are measured by expert-assessed visual scorings. Selection based on phenological characteristics is then conducted by investigating the coincidence of critical developmental stages with drought, heat, or other harsh environmental conditions (Sallam et al., 2019). Maulana et al. (2020) describes drought-related MTAs at the seedling stage of wheat (Table 3). In addition, drought stress during stem elongation can lead to yield reduction up to 71.52% (Ding et al., 2018). Early vigour, the rapid development of leaf area, has been genetically determined by 41 markers associated either with the NDVI or the projected leaf area, which could be used to select for varieties equipped with early vigour in the future (Vukasovic et al., 2022). Farhad et al. (2023) discovered several QTLs (i.e. QDtb.bisa.2D.4) that significantly relate to a shift in the time until booting (days to booting) towards earlier planting. MTAs on chromosomes 2B, 3A and 3D have been found to be related to the number of days to anthesis (Molero et al., 2019). Utilising genetics to select suitable varieties based on phenology is an important technique to face intense drought events. Understanding the link between genotype and phenology is essential to maximise grain yield in these scenarios.

Although these findings are significant and represent a substantial step towards crop optimisation against drought, there remains a large portion of heritability that is unaccounted for (see Section 4). As a result, multiple markers that could be useful for MAS might have gone unidentified. This missing heritability could be due to multiple testing correction or because the statistical tests assume a different distribution than that present in the actual data (Brachi et al., 2011), needing the development of new methods to tackle these issues.

6 Accelerating plant breeding processes with explainable AI

Artificial intelligence (AI) is now applied in many areas of the life sciences, thanks to the significant success of ML and particularly neural networks (NNs) as problem solvers (Holzinger et al., 2023a), which also has been enabled by the constant increase in computing power and resources. AI has already made its way into modern crop breeding, being used in the analysis of the increasing amount of plant image data, as well as in the modelling of GS and GWAS, overcoming some of the limitations of commonly used statistical methods (Harfouche et al., 2019; Jeon et al., 2023; Najafabadi et al., 2023). However, many AI algorithms have their caveats, as they often lack explainability and transparency due to their complex architecture. This is commonly referred to as the 'black box problem' (Castelvecchi, 2016), which can ultimately lead to the inability to provide users with explanations for their decisions. The emerging field of explainable AI (xAI) introduces new methods aiming to make AI systems more transparent and understandable (Arrieta et al., 2020; Miller et al., 2022; Holzinger et al., 2022b), laying the foundation for the digital transformation of smart agriculture, and especially plant breeding (Harfouche et al., 2019; Holzinger et al., 2022a).

6.1 Introduction into xAI methods

Although numerous xAI methods have been developed, and new ones continue to emerge for various NN architectures, no single xAI method or combination fully explains the decision-making process of the models. Each of them sheds light on a different aspect of the AI model's computation and many times it has been shown that there is no mutual consent between them, leading to the so-called 'disagreement' problem (Krishna et al., 2022). Currently, quality metrics for xAI methods (Doumard et al., 2023; Schwalbe and Finzel, 2023) and benchmarks for its evaluation are being defined (Agarwal et al., 2023) to motivate xAI research in directions that support trustworthy, reliable, actionable and causal explanations even if they don't always align with human pre-conceived notions and expectations (Holzinger et al., 2019; Magister et al., 2021; Finzel et al., 2022; Saranti et al., 2022; Cabitza et al., 2023; Holzinger et al., 2023c).

xAI methods have a coarse division between post-hoc and ante-hoc methods: the post-hoc ones are applied after the training has produced 'sufficiently' good results in terms of performance. For example, local interpretable model-agnostic explanation (LIME) (Ribeiro et al., 2016), which constructs local linear explanation models from the synthetic neighbourhood around the inputs, and Shapley additive explanations (SHAPs) (Shapley, 1952; Staniak and Biecek, 2019; Frye et al., 2020; Gevaert et al., 2023), which use game-theoretic notions to measure how influential features are to the prediction of a model, are procedures that could give scientists an interpretation of the 'black box' (Bach et al., 2015; Montavon et al., 2019; Amparore et al., 2021; den Broeck et al., 2022; Holzinger et al., 2022b). Counterfactual explanations, inspired by the work of Judea Pearl (Pearl and Mackenzie, 2018), are defined as all possibilities that deviate from the main course of events. In similar terms, the question 'what if' is applicable to counterfactual explanations that aim to provide information about features that, if they had different values, would result in a different output for the classification/regression problem (Sokol and Flach, 2019; Dandl et al., 2020). On the other hand, ante-hoc methods do not consist of individual software components applied after the model has converged and its internal parameters have solidified. Instead, they are models with built-in explainability. Decision trees (DTs) are one of the most representative models in this category and are widely used. They divide the space of possibilities into parts separated by feature ranges, making this method one of the easiest to understand (Safavian and Landgrebe, 1991). Generalised Additive Models go beyond linear and logistic regression, allowing the output to be expressed as an additive combination of pre-specified non-linear functions (Wood, 2004). Typically, the family of B-splines provides a balance between good performance and interpretability since these functions can be considered as individual and non-interacting. Bayesian Rule Lists contain IF-THEN statements in a list that describes the decision of the model (Letham et al., 2015). The Bayesian rule comes with the definition of a Dirichlet prior that specifies the number of pseudo-counts for a probability distribution, which is defined by a human domain expert (Koller and Friedman, 2009; Holzinger et al., 2023b). The posterior distribution is computed by a Bayesian update rule and

incorporation of the number of times one observed each output label.

Layer-wise relevance propagation (LRP) (Bach et al., 2015; Montavon et al., 2019) is a propagation-based method that uses the model's internal decision parameters to redistribute explanatory factors over the layers of the model, reaching the input variables and obtaining how important those are to the prediction and the model. While the computation of relevance of each feature or input component is something that is achieved by other methods, like sensitivity analysis (SA) (Simonyan et al., 2013) or SHAP, LRP uniquely computes both positive and negative relevance values. This is particularly important since the components that have positive relevance 'speak for' the result (e.g., the predicted class in a classification task), whereas those with negative relevance denote elements that contain evidence against the prediction and weaken the prediction confidence of the model. While this method is applied after the training of the model is accomplished, it is not entirely agnostic about the internal structure of the model. LRP has different variations for different NN architectures; for example, Long short-term memory (Hochreiter and Schmidhuber, 1997) networks have an adequately adapted LRP variation (Arras et al., 2017) that enables perturbation analysis of the input sequence and correspondingly graph neural networks (GNNs) have GNN-LRP (Schnake et al., 2020; Xiong et al., 2022) that uncovers positively and negatively important graph paths. LRP has been used for uncovering spurious correlations (so-called Clever-Hans phenomena) between the input and the output of an NN and also for clustered explanations with Whole Dataset Analysis (Lapuschkin et al., 2019).

6.2 Explainable AI methods for modern plant breeding

The plant breeding process, in its entirety, necessitates a high degree of transparency and explainability. Breeders, for instance, need more than just a predictive value to support their selection of genotypes; they rely on a wealth of information to understand the underlying biology and environmental interactions (Harfouche et al., 2019). xAI can be used to confer these qualities into effective ML models at several steps of the breeding process and in a multitude of ways:

- **Processing of UAV-sourced data:** AI is required to uncover the complex relationships between remotely acquired visual feedback and phenotypical traits (see Section 3.3). This is often a statistically ill-posed problem due to the challenges of replicating exact conditions from one year to another, the high number of external factors, and the cost of acquiring large-scale datasets of carefully measured phenotypical traits (Cheng et al., 2023). This statistical ambiguity can lead to both under- and over-fitting depending on the case. In this context, both ante-hoc and post-hoc xAI methods are important, as exemplified in Srivastava et al. (2022) for winter wheat yield prediction. Ante-hoc methods intervene in the form of strong regularisation, or inductive biases, which limit the space of possible models to those that closely follow a human-defined formulation of the problem. For example, Ge et al. (2023) predict rice distribution using a physically interpretable model trained directly using feature interpretation methods. These heavily regularised models often take the form of simple, interpretable algorithmic bricks that are trained to solve specific sub-problems, such as Tang et al. (2022) who integrate the domain knowledge that edge-detection is important directly into their winter wheat lodging detection architecture. Post-hoc methods serve as a necessary human-in-the-loop validation to counteract the difficulty to acquire enough data for a statistically significant validation. They serve as sanity checks that verify if the features deemed important by the model can be traced back to a physically understandable relationship. Such examples abound both inside (Sun et al., 2023) and outside (Temenos et al., 2022) winter wheat literature.
- **Understanding genotype-phenotype relationships:** AI can assist in unravelling the complex relationship between a plant's genotype and its phenotype in response to environmental conditions. Especially xAI can identify genetic variants that contribute to these traits, particularly those that have non-linear interactions - something that GWAS cannot do (Santorsola and Lescai, 2023). Feed-forward NNs go beyond association testing and can use several individuals with many SNPs to predict traits with an acceptable performance (Sharma et al., 2020). After the end of the training process, the xAI method DeepLift (Shrikumar et al., 2017) can be applied and computed for each input SNP attribution score that can take both positive and negative values (indicating the direction of contribution to the target variable). The SNPs with the highest attribution values can be thought of as potential causal causes and be investigated further for plausibility although the results of this research show that in cases of highly correlated features, DeepLift can perceive for one and the same model different input features (SNPs) as important. Building on their previous work (Mieth et al., 2016), Mieth et al. (2021) demonstrated that xAI can enhance traditional GWAS methods: NNs combined with statistical testing driven by xAI can provide a robust framework to uncover SNPs that play a decisive role in the classification result of the NN. LRP (Bach et al., 2015; Montavon et al., 2019) computes the relevance of each SNP used in the classification as if they were p values used to compute statistically significant associations. This approach surpasses the deficiencies of previous architectures that required Bonferroni correction for false rejections and returns additional as well as weak associations that might be significant. It also reduces the return of an incorrect association (statistical noise). However, the biological plausibility of the newly discovered SNPs needs to be validated, particularly if there are no existing GWAS results for them yet. Epistasis, the non-linear, non-

additive interaction between SNPs, is another important component of this relationship. It is often overlooked by classical GWAS methodologies, prompting the development of many techniques to try and dissect it (Niel et al., 2015). Romero (2022) describes an innovative process of extracting this behaviour from iterative RFs trained on this data (Basu et al., 2018). One of the advantages of using RF models is that their own architecture is easily interpretable Pfeifer et al. (2022).

- **Understanding complex interactions:** AI can be utilised for modelling and predicting how certain genotypes would react to conditions like drought. Unlike classical statistical multi-omics methods (Yazdani et al., 2022), AI is an effective tool for deciphering the complex interactions within a plant and those interactions a plant has with its environment, such as soil microbiome, weather, and other plants, which can influence its stress tolerance. xAI can provide insights into the reasoning behind these predictions, enhancing our understanding and facilitating targeted breeding strategies. In Niazian and Niedbała (2020), several cases of genotype-environment interactions (G × E) used by several AI models (having as input the genome sequence and output the phenotype) with their corresponding xAI methods were analysed, uncovering the decisive factors for these interactions (Streich et al., 2020). It is also shown that NNs outperform other AI models performance-wise on these tasks most of the time and the sensitivity analysis applied to the NNs detects the most important input variables for a prediction in different tasks such as assessment and classification of genetic diversity, yield component analysis and indirect selection (prediction), yield stability and G × E interaction, biotic and abiotic stress assessment, classical mating designs, and hybrid breeding programmes (Stein et al., 2022).

Scientific progress is based on understanding and explaining observable phenomena, and this is the advantage provided by the use of xAI. AI has been able to find complex relationships between genotype and phenotype, which could not have been found with other methods. However, it is important to apply these techniques with a higher degree of scientific rigour. Methods such as LRP, LIME, or SHAP are able to provide a deeper understanding of the behaviour of the model, and thus of the biological problem, a prerequisite in modern plant breeding (Harfouche et al., 2019).

7 Towards the implementation of modern tools for practical plant breeding

The previous sections have outlined the advantages of employing modern tools, such as GWAS for genetic marker characterisation, UAV-based remote sensing phenotyping, and the integration of xAI into the breeding process. In this concluding section, we aim to provide an overview of the practical tasks undertaken by plant breeders and how the aforementioned tools can improve the current state of the breeding process, as illustrated in Figure 4.

The responsibilities of a plant breeder can generally be categorised into the following steps: (i) defining the breeding objectives, (ii) creating genetic diversity, and (iii) selecting genotypes. Ultimately, a new variety is registered, certified seed is multiplied, and marketed (cf. Figure 4). The first step involves identifying key traits that will define a future variety. The second step aims to generate a high genetic diversity, particularly in target traits defined in Step 1, often with limited resources, such as a limited number of crosses or mutagenesis treatments. The third step is centred on the selection of candidate genotypes. This step is heavily reliant on data, necessitating efficient data collection and decision-making, often with limited (financial and human) resources, such as scorings, measurements, samplings and laboratory analysis, as well as downstream data analysis. Consequently, the methods and protocols developed by scientists often need to be scaled down or simplified for easy application within the breeding process.

For example, in GWAS, the ultimate objective is to develop markers taking advantage of a plethora of tools (cf. Table 2) with enough precision to predict the presence of a trait of interest. Eventually, these markers (cf. Table 3) should be utilised by the breeder, for instance, for screening potential crossing partners (Step 2) and MAS (Step 3). Saini et al. (2022) reviewed, that 86,122 wheat varieties have been analysed with GWAS, resulting in 46,940 loci for various agronomic, physiological, and quality traits. However, their implementation often remains a challenge in many breeding programmes due to several constraints, such as lack of transferability or additional disproportionate costs. Transferability concerns in GWAS are prevalent mostly between different populations and environments, as was shortly discussed in Section 4.2 (Guo et al., 2014; Blake et al., 2020; Mohammadi et al., 2020). Limited transferability due to relevant genotype by environment interactions can be addressed by, e.g., the inclusion of crop growth models (Technow et al., 2015). Mid-range genotyping platforms like KASP™ (Semagn et al., 2014) or MassArray® (Irwin, 2008) offer a relatively flexible, user-friendly, and affordable solution for practical breeding by being capable of screening tens to hundreds of markers in several hundreds of individuals. Both platforms have already been used to design ready-to-use assays for MAS in diverse sets of diploid crop species (e.g., Bomers et al., 2022), and have been successfully applied in polyploid wheat (Bérard et al., 2009; Rasheed et al., 2016; Makhoul et al., 2020; da Costa Lima Moraes et al., 2023; Liu et al., 2023) or aim to do so (Molin, 2024).

The objective of remote sensing phenotyping is to provide fast and precise phenotypic measurements. Particularly for plant breeding, UAV-based phenotyping offers an optimal combination of spatial resolution and speed of measurement (Figure 2). This data can be used by plant breeders primarily to enhance genotype selection (Step 3), but also to identify phenotypic diversity (Step 2). Numerous studies have applied UAV-based phenotyping in the context of plant breeding in the past (White et al., 2012; Chapman et al., 2014; Araus et al., 2018; Thenkabail et al., 2018), using a variety of sensors including multi- and hyperspectral, thermal,

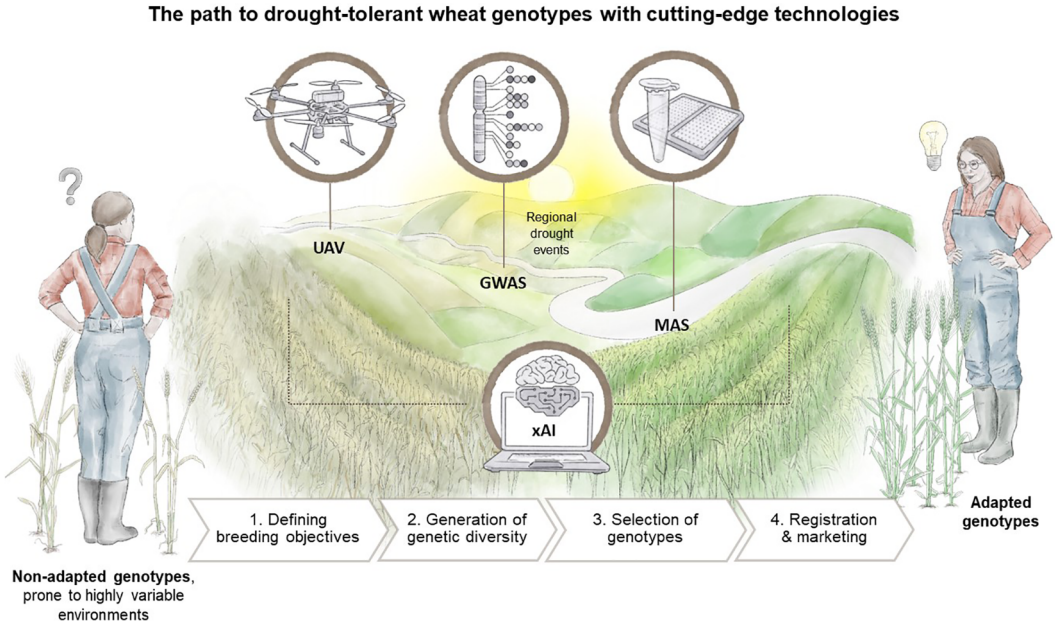


FIGURE 4
The road to drought-tolerant wheat genotypes remains tedious and time-consuming, but cutting-edge technologies promise to speed up the breeding process. UAV-based phenotyping, GWAS, and ultimately MAS are being increasingly used. However, a cross-process and crucial role is played by xAI, which is not only applied in data analysis and interpretation, but also in decision-making within the entire breeding process, from the definition of breeding goals to the final step of registration and marketing. Designed by Tatjana Hirschmugl.

RGB, and LiDAR to investigate traits such as yield, biomass, plant height, crop health and stress, diseases, pests, as well as nutrient and water content (Yang et al., 2009; Prabhakar et al., 2012; Virnodkar et al., 2020; Zhou et al., 2021b; Thoday-Kennedy et al., 2022; Hütt et al., 2023; Joshi et al., 2023). However, its application in practical breeding is still limited (Matese et al., 2023) due to the need for expertise in several areas, such as drone piloting, legislation, flight planning, photogrammetric processing as well as data processing, modelling, and analysis (White et al., 2012; Chapman et al., 2014; Reynolds et al., 2020; Guo et al., 2021).

In the current scientific dialogue, AI has emerged as a vital tool for problem-solving and knowledge discovery, particularly in the life sciences (Holzinger et al., 2023a). Its applications are manifold and extend to specialised fields like plant breeding. In this context, AI facilitates the analysis of plant image data and plays a crucial role in GWAS and genomic selection (Zhang et al., 2017; Parmley et al., 2019; Aono et al., 2022). AI's usefulness extends beyond data analysis and permeates the entire decision-making pipeline as depicted in Figure 4, from initial data collection and preprocessing (step 1), to feature selection and modelling (step 2), and finally to evaluation and interpretation of results (step 3). The technology's versatility and computational prowess allow it to process large datasets, discern patterns that may be overlooked by human experts, and provide actionable insights. Essentially, AI acts as a decision support system that enhances the abilities of domain specialists, such as plant breeders, by furnishing them with more accurate and comprehensive information.

The emergence of xAI further enhances the utility of AI in plant breeding. xAI aims to make the complex decision-making processes of AI algorithms transparent and understandable. This is achieved through various methods, such as feature importance ranking, DTs, and counterfactual explanations, among others Holzinger et al. (2021). The increased transparency provided by xAI not only unravels the black-box nature of complex algorithms but also promotes trust and acceptance among human decision-makers. The importance of xAI goes beyond mere understanding of AI's operations; it addresses ethical and accountability concerns by ensuring that algorithmic decisions can be audited and justified Müller et al. (2022). This is particularly important in high-stakes applications like plant breeding, where decisions can have enduring impacts on agricultural productivity and sustainability. Therefore, the integration of xAI into decision-making processes enhances the trustworthiness and acceptance of AI systems, paving the way for more responsible and effective applications of AI in the life sciences (Holzinger et al., 2022a), including specialised domains such as plant breeding (Harfouche et al., 2019).

In summary, the cutting-edge tools reviewed in this study, encompassing UAV-based phenotyping, GWAS, MAS, bolstered by ML, and the integration of xAI, collectively represent a transformative shift in plant breeding (Figure 4). These innovative methods have the potential to revolutionise the way how breeders gather field data, interpret it, and ultimately make informed decisions throughout the entire breeding process, representing a new era in smart agriculture. By leveraging these technological capabilities, breeders can significantly accelerate the development of new crop varieties with improved traits, such as drought tolerance. This acceleration not only reflects the progress in science and technology but also holds the promise of addressing critical agricultural challenges, such as feeding an expanding global population and mitigating the effects of climate change on crop production.

Author contributions

IC-B: Writing – original draft, Writing – review & editing, Investigation. LJK: Conceptualization, Writing – original draft, Writing – review & editing, Investigation. LB: Writing – original draft, Conceptualization, Investigation, Writing – review & editing. GB: Writing – original draft, Writing – review & editing. AS: Writing – original draft, Writing – review & editing. JS: Writing – review & editing, Writing – original draft. PF-J: Investigation, Supervision, Writing – review & editing, Conceptualization. CS: Supervision, Writing – review & editing. FB: Writing – review & editing. FT: Writing – original draft, Writing – review & editing. MS-Z: Investigation, Writing – review & editing. EZ: Supervision, Writing – review & editing. AH: Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. EMM: Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This review has been conducted in the frame of the d4agrotech initiative (www.d4agrotech.at), in which the project 'WheatVIZ' (WST3-F-5030665/018-2022) received funding from the Federal Government of Lower Austria (Grantholder: EMM). Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 'explainable Artificial Intelligence' (Grantholder: AH).

Acknowledgments

With regard to the data underlying Figure 1A, we acknowledge the World Climate Research Programme, which coordinated and promoted CMIP6, and we thank the climate modelling groups for producing and making available their model output, and the Earth System Grid Federation for archiving the data and providing access. A big thank you goes to Tatjana Hirschmugl (www.scillustration.at) for the design of all illustrations. Many thanks also go to Philipp Karoshi for providing expert knowledge.

Conflict of interest

Authors LJK and MS-Z are employed by the company Saatzucht Edelhof GmbH.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Abdalla, M., Ahmed, M. A., Cai, G., Wankmüller, F., Schwartz, N., Litig, O., et al. (2022). Stomatal closure during water deficit is controlled by below-ground hydraulics. *Ann. Bot.* 129, 161–170. doi: 10.1093/aob/mcab141

Adão, T., Hruška, J., Pádua, L., Bessa, J., Peres, E., Morais, R., et al. (2017). Hyperspectral imaging: A review on uav-based sensors, data processing and applications for agriculture and forestry. *mdpi.com* 9, 1110. doi: 10.3390/rs9111110

Affortit, P., Ahmed, M. A., Grondin, A., Delzon, S., Carminati, A., and Laplaze, L. (2023). Keep in touch: the soil–root hydraulic continuum and its role in drought resistance in crops. *J. Exp. Bot.* 75 (2), 584–593. doi: 10.1093/jxb/erad312

Agarwal, C., Queen, O., Lakkaraju, H., and Zitnik, M. (2023). Evaluating explainability for graph neural networks. *Sci. Data* 10, 144. doi: 10.1038/s41597-023-01974-x

AgEagle Aerial Systems Inc (2023) *Altum-pt - drone sensors*. Available online at: https://ageagle.com/drone-sensors/altumpt-camera/.

AGES (2023). *Österreichische Beschreibende Sortenliste 2023 Landwirtschaftliche Pflanzenarten*, Schriftenreihe 21/2023. Vienna, Austria: Österreichische Agentur für Gesundheit und Ernährungssicherheit GmbH.

Ahmed, A. A., Mohamed, E. A., Hussein, M. Y., and Sallam, A. (2021). Genomic regions associated with leaf wilting traits under drought stress in spring wheat at the seedling stage revealed by gwas. *Environ. Exp. Bot.* 184, 104393. doi: 10.1016/j.envexpbot.2021.104393

Akram, S., Arif, M. A. R., and Hameed, A. (2021). A gbs-based gwas analysis of adaptability and yield traits in bread wheat (triticum aestivum l.). *J. Appl. Genet.* 62, 27–41. doi: 10.1007/s13353-020-00593-1

Alamin, M., Sultana, M. H., Lou, X., Jin, W., and Xu, H. (2022). Dissecting complex traits using omics data: A review on the linear mixed models and their application in gwas. *Plants* 11, 3277. doi: 10.3390/plants11233277

Allard, R. W., and Bradshaw, A. D. (1964). Implications of genotype-environmental interactions in applied plant breeding. *Crop Sci.* 4, 503–508. doi: 10.2135/cropsci1964.0011183X000400050021x

Aloisi, I., Yacoubi, I., Gadaleta, A., Schwember, A. R., and Marcotuli, I. (2023). Editorial: Exploiting wheat biodiversity and agricultural practices for tackling the effects of climate change. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1257502

Amparore, E., Perotti, A., and Bajardi, P. (2021). To trust or not to trust an explanation: using leaf to evaluate local linear xai methods. *PeerJ Comput. Sci.* 7, e479. doi: 10.7717/peerj-cs.479

Anderegg, J., Yu, K., Aasen, H., Walter, A., Liebisch, F., and Hund, A. (2020). Spectral vegetation indices to track senescence dynamics in diverse wheat germplasm. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01749

Ang, K. L.-M., and Seng, J. K. P. (2021). Big data and machine learning with hyperspectral information in agriculture. *IEEE Access* 9, 36699–36718. doi: 10.1109/Access.6287639

>Ansarifar, J., Wang, L., and Archontoulis, S. V. (2021). An interaction regression model for crop yield prediction. *Sci. Rep.* 11, 17754. doi: 10.1038/s41598-021-97221-7

Aono, A. H., Ferreira, R. C. U., da Costa Lima Moraes, A., de Castro Lara, L. A., Pimenta, R. J. G., Costa, E. A., et al. (2022). A joint learning approach for genomic prediction in polyploid grasses. *Sci. Rep.* 12, 12499. doi: 10.1038/s41598-022-16417-7

Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, eaar7191. doi: 10.1126/science.aar7191

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., and Cairns, J. E. (2018). Translating high-throughput phenotyping into genetic gain. *Trends Plant Sci.* 23, 451–466. doi: 10.1016/j.tplants.2018.02.001

Araus, J. L., Slafer, G. A., Royo, C., and Serret, M. D. (2008). Breeding for yield potential and stress adaptation in cereals. *Crit. Rev. Plant Sci.* 27, 377–412. doi: 10.1080/07352680802467736

Arras, L., Montavon, G., Müller, K. R., and Samek, W. (2017). "Explaining recurrent neural network predictions in sentiment analysis," in *EMNLP 2017 - 8th Workshop on*

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA 2017 - Proceedings of the Workshop. 159–168. doi: 10.18653/v1/w17-5221

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Inf. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Arruda, M., Lipka, A., Brown, P., Krill, A., Thurber, C., Brown-Guedira, G., et al. (2016). Comparing genomic selection and marker-assisted selection for fusarium head blight resistance in wheat (triticum aestivum l.). *Mol. Breed.* 36, 1–11. doi: 10.1007/s11032-016-0508-5

Atkinson, J. A., Jackson, R. J., Bentley, A. R., Ober, E., and Wells, D. M. (2018). *Field Phenotyping for the Future* (Nottingham, UK: Wiley). doi: 10.1002/9781119312994.apr0651

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* 10, e0130140. doi: 10.1371/journal.pone.0130140

Bai, G., Ge, Y., Hussain, W., Baenziger, P. S., and Graef, G. (2016). A multi-sensor system for high throughput field phenotyping in soybean and wheat breeding. *Comput. Electron. Agric.* 128, 181–192. doi: 10.1016/j.compag.2016.08.021

Ball, R. D. (2013). Designing a gwas: Power, sample size, and data structure. *Methods Mol. Biol.* 1019, 37–98. doi: 10.1007/978-1-62703-447-03/COVER

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci.* 115, 1943–1948. doi: 10.1073/pnas.1711236115

Becker, S. R., Byrne, P. F., Reid, S. D., Bauerle, W. L., McKay, J. K., and Haley, S. D. (2016). Root traits contributing to drought tolerance of synthetic hexaploid wheat in a greenhouse study. *Euphytica* 207, 213–224. doi: 10.1007/s10681-015-1574-1

Becker, E., and Schmidhalter, U. (2017). Evaluation of yield and drought using active and passive spectral sensing systems at the reproductive stage in wheat. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00379

Beltrame, L., Salzinger, J., Fanta-Jende, P., and Sulzbachner, C. (2024). "Practical strategies for automated phenotyping: from raw UAV data to multispectral time series for machine learning applications," in Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs (Ed), *74. Jahrestagung 2023, 20-22 November, Raumberg-Gumpenstein*. (Vienna, Austria: *University of Natural Resources and Life Sciences*), 5–10.

Bennani, S., Birouk, A., Jlibene, M., Sanchez-Garcia, M., Nsarellah, N., Gaboun, F., et al. (2022). Drought-tolerance qtls associated with grain yield and related traits in spring bread wheat. *Plants* 11, 986. doi: 10.3390/plants11070986

Bérard, A., Paslier, M. C. L., Dardevet, M., Exbrayat-Vinson, F., Bonnin, I., Cenci, A., et al. (2009). High-throughput single nucleotide polymorphism genotyping in wheat (¡i¿triticum¡/i¿ spp.). *Plant Biotechnol. J.* 7, 364–374. doi: 10.1111/j.1467-7652.2009.00404.x

Bhandari, M., Ibrahim, A. M., Xue, Q., Jung, J., Chang, A., Rudd, J. C., et al. (2020). Assessing winter wheat foliage disease severity using aerial imagery acquired from small unmanned aerial vehicle (uav). *Comput. Electron. Agric.* 176, 105665. doi: 10.1016/j.compag.2020.105665

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. (2018). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *Int. J. Mol. Sci.* 19, 3011. doi: 10.3390/ijms19103011

Bilgrami, S. S., Ramandi, H. D., Shariati, V., Razavi, K., Tavakol, E., Fakheri, B. A., et al. (2020). Detection of genomic regions associated with tiller number in Iranian bread wheat under different water regimes using genome-wide association study. *Sci. Rep.* 10, 14034. doi: 10.1038/s41598-020-69442-9

Blake, R., Gina, B.-G., H., S. C., and Mohsen, M. (2020). Transferability of marker trait associations in wheat is disturbed mainly by genotype × year interaction. *Crop Breeding Genet. Genomics*. 4. doi: 10.20900/cbgg20200013

Blum, A. (2009). Effective use of water (euw) and not water-use efficiency (wue) is the target of crop yield improvement under drought stress. *Field Crops Res.* 112, 119–123. doi: 10.1016/j.fcr.2009.03.009

Blum, A. (2011). *Plant Breeding for Water-Limited Environments* (Amsterdam, Netherlands: Springer New York). doi: 10.1007/978-1-4419-7491-4

Bodner, G., Nakhforoosh, A., and Kaul, H.-P. (2015). Management of crop water under drought: a review. *Agron. Sustain. Dev.* 35, 401–442. doi: 10.1007/s13593-015-0283-4

Bogard, M., Hourcade, D., Piquemal, B., Gouache, D., Deswartes, J.-C., Throude, M., et al. (2021). Marker-based crop model-assisted ideotype design to improve avoidance of abiotic stress in bread wheat. *J. Exp. Bot.* 72, 1085–1103. doi: 10.1093/jxb/eraa477

Bomers, S., Sehr, E. M., Adam, E., von Gehren, P., Hansel-Hohl, K., Prat, N., et al. (2022). Towards heat tolerant runner bean (phaseolus coccineus l.) by utilizing plant genetic resources. *Agronomy* 12, 612. doi: 10.3390/agronomy12030612

Bouaziz, M., Ambroise, C., and Guedj, M. (2011). Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *PloS One* 6, e28845. doi: 10.1371/journal.pone.0028845

Bourrat, P., Lu, Q., and Jablonka, E. (2017). Why the missing heritability might not be in the dna. *BioEssays* 39, 1700067. doi: 10.1002/bies.201700067

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Brachi, B., Morris, G. P., and Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol.* 12, 232. doi: 10.1186/gb-2011-12-10-232

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). Tassel: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F. X., and Huard, F. (2010). Why are wheat yields stagnating in europe? a comprehensive data analysis for France. *Field Crops Res.* 119, 201–212. doi: 10.1016/j.fcr.2010.07.012

Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K., Kumar, A., Pradhan, A. K., et al. (2022). Genomic selection: a tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Front. Genet.* 13. doi: 10.3389/fgene.2022.832153

Buerstmayr, H., Steiner, B., Lemmens, M., and Ruckenbauer, P. (2000). Resistance to fusarium head blight in winter wheat: Heritability and trait associations. *Crop Sci.* 40, 1012–1018. doi: 10.2135/cropsci2000.4041012x

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., et al. (2023). Quod erat demonstrandum? - towards a typology of the concept of explanation for the design of explainable ai. *Expert Syst. Appl.* 213, 118888. doi: 10.1016/j.eswa.2022.118888

Carlson, T. N., and Ripley, D. A. (1997). On the relation between ndvi, fractional vegetation cover, and leaf area index. *Remote Sens. Environ.* 62, 241–252. doi: 10.1016/S0034-4257(97)00104-1

Carminati, A., and Vetterlein, D. (2013). Plasticity of rhizosphere hydraulic properties as a key for efficient utilization of scarce resources. *Ann. Bot.* 112, 277–290. doi: 10.1093/aob/mcs262

Castelvecchi, D. (2016). Can we open the black box of ai? *Nature* 538, 20–23. doi: 10.1038/538020a

Cattivelli, L., Rizza, F., Badeck, F. W., Mazzucotelli, E., Mastrangelo, A. M., Francia, E., et al. (2008). Drought tolerance improvement in crop plants: An integrated view from breeding to genomics. *Field Crops Res.* 105, 1–14. doi: 10.1016/j.fcr.2007.07.004

Chandel, N. S., Chakraborty, S. K., Rajwade, Y. A., Dubey, K., Tiwari, M. K., and Jat, D. (2021). Identifying crop water stress using deep learning models. *Neural Computing Appl.* 33, 5353–5367. doi: 10.1007/s00521-020-05325-4

Chang, T.-G., Chang, S., Song, Q.-F., Perveen, S., and Zhu, X.-G. (2019). Systems models, phenomics and genomics: three pillars for developing high-yielding photosynthetically efficient crops. *silico Plants* 1, diy003. doi: 10.1093/insilicoplants/diy003

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Secondgeneration plink: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi: 10.1186/s13742-015-0047-8

Chang, T.-G., and Zhu, X.-G. (2017). Source–sink interaction: a century old concept under the light of modern molecular systems biology. *J. Exp. Bot.* 68, 4417–4431. doi: 10.1093/jxb/erx002

Chapman, S., Merz, T., Chan, A., Jackway, P., Hrabar, S., Dreccer, M., et al. (2014). Pheno-copter: A low-altitude, autonomous remote-sensing robotic helicopter for high-throughput field-based phenotyping. *Agronomy* 4, 279–301. doi: 10.3390/agronomy4020279

Chawade, A., van Ham, J., Blomquist, H., Bagge, O., Alexandersson, E., and Ortiz, R. (2019). Highthroughput field-phenotyping tools for plant breeding and precision agriculture. *Agronomy* 9, 258. doi: 10.3390/agronomy9050258

Cheng, X., Sun, Y., Zhang, W., Wang, Y., Cao, X., and Wang, Y. (2023). Application of deep learning in multitemporal remote sensing image classification. *Remote Sens.* 15, 3859. doi: 10.3390/rs15194705

Clyde, D. (2019). Making the case for more inclusive gwas. *Nat. Rev. Genet.* 20, 500–501. doi: 10.1038/s41576-019-0160-0

Collard, B. C., and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc. B: Biol. Sci.* 363, 557–572. doi: 10.1098/rstb.2007.2170

Collins, B., Chapman, S., Hammer, G., and Chenu, K. (2021). Limiting transpiration rate in high evaporative demand conditions to improve Australian wheat productivity. *silico Plants* 3, diab006. doi: 10.1093/insilicoplants/diab006

Condorelli, G. E., Maccaferri, M., Newcomb, M., Andrade-Sanchez, P., White, J. W., French, A. N., et al. (2018). Comparative aerial and ground based high throughput phenotyping for the genetic dissection of ndvi as a proxy for drought adaptive traits in durum wheat. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00893

Cortes, L. T., Zhang, Z., and Yu, J. (2021). Status and prospects of genome-wide association studies in plants. *Plant Genome* 14, e20077. doi: 10.1002/tpg2.20077

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the r package sommer. *PloS One* 11, e0156744. doi: 10.1371/journal.pone.0156744

Cruiziat, P., Cochard, H., and Améglio, T. (2002). Hydraulic architecture of trees: main concepts and results. *Ann. For. Sci.* 59, 723–752. doi: 10.1051/forest:2002060

Cummings, C., Miao, Y., Paiao, G. D., Kang, S., and Fernández, F. G. (2021). Corn nitrogen status diagnosis with an innovative multi-parameter crop circle phenom sensing system. *Remote Sens.* 13, 401. doi: 10.3390/rs13030401

Czyż, E. A., Dexter, A. R., Czyż, E. A., and Dexter, A. R. (2012). Plant wilting can be caused either by the plant or by the soil. *Soil Res.* 50, 708–713. doi: 10.1071/SR12189

Daba, S. D., Tyagi, P., Brown-Guedira, G., and Mohammadi, M. (2018). Genome-wide association studies to identify loci and candidate genes controlling kernel weight and length in a historical United States wheat population. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01045

da Costa Lima Moraes, A., Sforc¸a, D. A., Mancini, M. C., Vigna, B. B. Z., and de Souza, A. P. (2023). Polyploid snp genotyping using the massarray system. *Methods Mol. Biol. (Clifton N.J.)* 2638, 93–113. doi: 10.1007/978-1-0716-3024-27/COVER

Dallinger, H. G., Löschenberger, F., Azrak, N., Ametz, C., Michel, S., and Bürstmayr, H. (2023). Genome-wide association mapping for pre-harvest sprouting in european winter wheat detects novel resistance qtl, pleiotropic effects, and structural variation in multiple genomes. *Plant Genome* e20301. doi: 10.1002/tpg2.20301

Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). "Multi-objective counterfactual explanations," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (Cham, Switzerland: Springer International Publishing), 12269, 448–469. doi: 10.1007/978-3-030-58112-131/FIGURES/6

Darra, N., Anastasiou, E., Kriezi, O., Lazarou, E., Kalivas, D., and Fountas, S. (2023). Can yield prediction be fully digitilized? a systematic review. *Agronomy* 13, 2441. doi: 10.3390/agronomy13092441

Das, S., Christopher, J., Apan, A., Choudhury, M. R., Chapman, S., Menzies, N. W., et al. (2021). Evaluation of water status of wheat genotypes to aid prediction of yield on sodic soils using uav-thermal imaging and machine learning. *Agric. For. Meteorology* 307, 108477. doi: 10.1016/j.agrformet.2021.108477

Davidson, H. R., and Campbell, C. A. (2011). Growth rates, harvest index and moisture use of manitou spring wheat as influenced by nitrogen, temperature and moisture. *Can. J. Plant Sci.* 64, 825–839. doi: 10.4141/CJPS84-114

Degen, B., and Müller, N. (2023). *Advanced marker-assisted selection versus genomic selection in breeding programs* (Cold Spring Harbor Laboratory eprint). Available at: https://www.biorxiv.org/content/early/2023/02/22/2023.02.20.529263.full.pdf.

den Broeck, G. V., Lykov, A., Schleich, M., and Suciu, D. (2022). On the tractability of shap explanations. *J. Artif. Intell. Res.* 74, 851–886. doi: 10.1613/jair.1.13283

Devate, N. B., Krishna, H., Parmeshwarappa, S. K. V., Manjunath, K. K., Chauhan, D., Singh, S., et al. (2022). Genome-wide association mapping for component traits of drought and heat tolerance in wheat. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.943033

Ding, J., Huang, Z., Zhu, M., Li, C., Zhu, X., and Guo, W. (2018). Does cyclic water stress damage wheat yield more than a single stress? *PloS One* 13, e0195535. doi: 10.1371/JOURNAL.PONE.0195535

Doumard, E., Aligon, J., Escriva, E., Excoffier, J. B., Monsarrat, P., and Soulé-Dupuy, C. (2023). A quantitative approach for the comparison of additive local explanation methods. *Inf. Syst.* 114, 102162. doi: 10.1016/j.is.2022.102162

Duan, T., Chapman, S. C., Guo, Y., and Zheng, B. (2017). Dynamic monitoring of ndvi in wheat agronomy and breeding trials using an unmanned aerial vehicle. *Field Crops Res.* 210, 71–80. doi: 10.1016/j.fcr.2017.05.025

Eitel, J. U., Long, D. S., Gessler, P. E., and Smith, A. M. (2007). Using *in-situ* measurements to evaluate the new rapideye™ satellite series for prediction of wheat nitrogen status. *Int. J. Remote Sens.* 28, 4183–4190. doi: 10.1080/01431160701422213

Enoma, D. O., Bishung, J., Abiodun, T., Ogunlana, O., and Osamor, V. C. (2022). Machine learning approaches to genome-wide association studies. *J. King Saud Univ. - Sci.* 34, 101847. doi: 10.1016/j.jksus.2022.101847

European Space Agency (2023a) *Sentinel-2*. Available online at: https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

European Space Agency (2023b) *Worldview-3*. Available online at: https://earth.esa.int/eogateway/missions/worldview-3.

European Space Agency (2024) *Rapideye dataset*. Available online at: https://earth.esa.int/eogateway/missions/rapideyedatasection.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., et al. (2016). Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model. Dev.* 9, 1937–1958. doi: 10.5194/gmd-9-1937-2016

Fang, H., Baret, F., Plummer, S., and Schaepman-Strub, G. (2019). An overview of global leaf area index (lai): Methods, products, validation, and applications. *Rev. Geophysics* 57, 739–799. doi: 10.1029/2018RG000608

FAO (2010). *Aquastat* (Rome: Food and Agriculture Organization).

FAO (2023) *Faostat. License: CC BY-NC-SA 3.0 IGO*. Available online at: https://www.fao.org/ (Accessed 07-06-2023).

Farhad, M., Tripathi, S., Singh, R., Joshi, A., Bhati, P., Vishwakarma, M., et al. (2023). Gwas for early-establishment qtls and their linkage to major phenology-affecting genes (vrn, ppd, and eps) in bread wheat. *Genes* 14, 1507. doi: 10.3390/genes14071507

Fei, S., Hassan, M. A., Xiao, Y., Su, X., Chen, Z., Cheng, Q., et al. (2023). Uav-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precis. Agric.* 24, 187–212. doi: 10.1007/s11119-022-09938-8

Fernandez-Gallego, J. A., Lootens, P., Borra-Serrano, I., Derycke, V., Haesaert, G., Roldán-Ruiz, I., et al. (2020). Automatic wheat ear counting using machine learning based on rgb uav imagery. *Plant J.* 103, 1603–1613. doi: 10.1111/tpj.14799

Fernando, R. L., and Garrick, D. (2013). Bayesian methods applied to gwas. *Methods Mol. Biol.* 1019, 237–274. doi: 10.1007/978-1-62703-447-010/COVER

Finzel, B., Saranti, A., Angerschmid, A., Tafler, D., Pfeifer, B., and Holzinger, A. (2022). Generating explanations for conceptual validation of graph neural networks: An investigation of symbolic predicates learned on relevance-ranked sub-graphs. *KI - Kunstliche Intelligenz* 36, 271–285. doi: 10.1007/s13218-022-00781-7

Fitzgerald, G., Rodriguez, D., and O'Leary, G. (2010). Measuring and predicting canopy nitrogen nutrition in wheat using a spectral index—the canopy chlorophyll content index (ccci). *Field Crops Res.* 116, 318–324. doi: 10.1016/j.fcr.2010.01.010

Frye, C., Rowat, C., and Feige, I. (2020). Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Adv. Neural Inf. Process. Syst.* 33, 1229–1239. doi: 10.48550/arXiv.1910.06358

Furbank, R. T., and Tester, M. (2011). Phenomics - technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* 16, 635–644. doi: 10.1016/j.tplants.2011.09.005

Gahlaut, V., Jaiswal, V., Balyan, H. S., Joshi, A. K., and Gupta, P. K. (2021). Multi-locus gwas for grain weight-related traits under rain-fed conditions in common wheat (triticum aestivum l.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.758631

Galieni, A., D'Ascenzo, N., Stagnari, F., Pagnani, G., Xie, Q., and Pisante, M. (2021). Past and future of plant stress detection: An overview from remote sensing to positron emission tomography. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.609155

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-gwas era: From association to function. *Am. J. Hum. Genet.* 102, 717–730. doi: 10.1016/j.ajhg.2018.04.002

Galluzzi, G., Seyoum, A., Halewood, M., Noriega, I. L., and Welch, E. W. (2020). The role of genetic resources in breeding for climate change: The case of public breeding programmes in eighteen developing countries. *Plants* 9, 1129. doi: 10.3390/plants9091129

Gao, B. C. (1996). Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* 58, 257–266. doi: 10.1016/S0034-4257(96)00067-3

Gao, Z., Bian, J., Lu, F., Jiao, Y., and He, H. (2023). Triticeae crop genome biology: an endless frontier. *Front. Plant Sci.* 14. doi: 10.3389/FPLS.2023.1222681

Gaut, B. S., and Long, A. D. (2003). The lowdown on linkage disequilibrium. *Plant Cell* 15, 1502–1506. doi: 10.1105/tpc.150730

Gautam, T., Amardeep,, Saripalli, G., Rakhi,, Kumar, A., Gahlaut, V., et al. (2021). Introgression of a drought insensitive grain yield qtl for improvement of four Indian bread wheat cultivars using marker assisted breeding without background selection. *J. Plant Biochem. Biotechnol.* 30, 172–183. doi: 10.1007/s13562-020-00553-0

Ge, J., Zhang, H., Xu, L., Sun, C., Duan, H., Guo, Z., et al. (2023). A physically interpretable rice field extraction model for polsar imagery. *Remote Sens.* 15, 974. doi: 10.3390/rs15040974

Geisler, G. (1983). *Ertragsphysiologie von Kulturarten des Gemässigten Klimas* (Berlin and Hamburg, Germany: Parey).

GEO3D (2023) *Open titan*. Available online at: https://www.geo3d.hr/3d-laser-scanners/teledyne-optech/optech-titan.

Gevaert, A., Saranti, A., Holzinger, A., and Saeys, Y. (2023). "Efficient approximation of asymmetric shapley values using functional decomposition," in *Machine Learning and Knowledge Extraction*. Eds. A. Holzinger, P. Kieseberg, F. Cabitza, A. Campagner, A. M. Tjoa and E. Weippl (Springer Nature Switzerland, Cham), 13–30. doi: 10.1007/978-3-031-40837-32

Gilbert, N. (2010). How to avert a global water crisis. *Nature*. doi: 10.1038/news.2010.490

Giménez-Gallego, J., González-Teruel, J. D., Jiménez-Buendía, M., Toledo-Moreo, A. B., Soto-Valles, F., and Torres-Sánchez, R. (2019). Segmentation of multiple tree leaves pictures with natural backgrounds using deep learning for image-based agriculture applications. *Appl. Sci.* 10, 202. doi: 10.3390/APP10010202

Gitelson, A. A., Merzlyak, M. N., and Chivkunova, O. B. (2001). Optical properties and nondestructive estimation of anthocyanin content in plant leaves. *Photochem. Photobiol.* 74, 38–45. doi: 10.1562/0031-8655(2001)074<0038:OPANEO>2.0.CO;2

Gizaw, S. A., Godoy, J. G. V., Garland-Campbell, K., and Carter, A. H. (2018). Genome-wide association study of yield and component traits in pacific northwest winter wheat. *Crop Sci.* 58, 2315–2330. doi: 10.2135/cropsci2017.12.0740

Gogna, A., Zhang, J., Jiang, Y., Schulthess, A. W., Zhao, Y., and Reif, J. C. (2023). Filtering for snps with high selective constraint augments mid-parent heterosis predictions in wheat (triticum aestivum l.). *Crop J.* 11, 166–176. doi: 10.1016/j.cj.2022.06.009

González-Navarro, O. E., Griffiths, S., Molero, G., Reynolds, M. P., and Slafer, G. A. (2015). Dynamics of floret development determining differences in spike fertility in an elite population of wheat. *Field Crops Res.* 172, 21–31. doi: 10.1016/j.fcr.2014.12.001

Grinberg, N. F., Orhobor, O. I., and King, R. D. (2020). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Mach. Learn.* 109, 251–277. doi: 10.1007/s10994-019-05848-5

Gu, J., Yin, X., Stomph, T.-J., and Struik, P. C. (2014). Can exploiting natural genetic variation in leaf photosynthesis contribute to increasing rice productivity? a simulation analysis. *Plant Cell Environ.* 37, 22–34. doi: 10.1111/pce.12173

Guan, J., Garcia, D. F., Zhou, Y., Appels, R., Li, A., and Mao, L. (2020). The battle to sequence the bread wheat genome: A tale of the three kingdoms. *Genomics Proteomics Bioinf.* 18, 221–229. doi: 10.1016/j.gpb.2019.09.005

Guo, W., Carroll, M. E., Singh, A., Swetnam, T. L., Merchant, N., Sarkar, S., et al. (2021). Uas-based plant phenotyping for research and breeding applications. *Plant Phenomics* 2021, e20077. doi: 10.34133/2021/9840192

Guo, Z., Tucker, D. M., Basten, C. J., Gandhi, H., Ersoz, E., Guo, B., et al. (2014). The impact of population structure on genomic prediction in stratified populations. *Theor. Appl. Genet.* 127, 749–762. doi: 10.1007/s00122-013-2255-x

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinf.* 12, 1–12. doi: 10.1186/1471-2105-12-186/FIGURES/2

Haboudane, D., Miller, J. R., Tremblay, N., Zarco-Tejada, P. J., and Dextraze, L. (2002). Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sens. Environ.* 81, 416–426. doi: 10.1016/S0034-4257(02)00018-4

Hack, H., Bleiholder, H., Buhr, L., Meier, U., Schnock-Fricke, U., Weber, E., et al. (1992). Uniform coding of the phenological developmental stages of monocotyledonous and dicotyledonous plants, bbch scale, general. *News Sheet German Plant Prot. Service* 44, 265–270.

Haghighattalab, A., Pérez, L. G., Mondal, S., Singh, D., Schinstock, D., Rutkoski, J., et al. (2016). Application of unmanned aerial systems for high throughput phenotyping of large wheat breeding nurseries. *Plant Methods* 12, 1–15. doi: 10.1186/s13007-016-0134-6

Hakula, A., Ruoppa, L., Lehtomäki, M., Yu, X., Kukko, A., Kaartinen, H., et al. (2023). Individual tree segmentation and species classification using high-density close-range multispectral laser scanning data. *ISPRS Open J. Photogrammetry Remote Sens.* 9, 100039. doi: 10.1016/j.ophoto.2023.100039

Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., et al. (2006). Models for navigating biological complexity in breeding improved crop plants. *Trends Plant Sci.* 11, 587–593. doi: 10.1016/j.tplants.2006.10.006

Hammer, P., and Hopper, D. (1997). *Experimental design* (Cambridge, Massachusetts, USA: Iowa State University Ames).

Harfouche, A. L., Jacobson, D. A., Kainer, D., Romero, J. C., Harfouche, A. H., Mugnozza, G. S., et al. (2019). Accelerating climate resilient plant breeding by applying next-generation artificial intelligence. *Trends Biotechnol.* 37, 1217–1235. doi: 10.1016/j.tibtech.2019.05.007

Hassan, M. A., Yang, M., Rasheed, A., Yang, G., Reynolds, M., Xia, X., et al. (2019). A rapid monitoring of ndvi across the wheat growth cycle for grain yield prediction using a multi-spectral uav platform. *Plant Sci.* 282, 95–103. doi: 10.1016/j.plantsci.2018.10.022

Hazelett, D. J., Conti, D. V., Han, Y., Olama, A. A. A., Easton, D., Eeles, R. A., et al. (2016). Reducing gwas complexity. *Cell Cycle* 15, 22–24. doi: 10.1080/15384101.2015.1120928

Heino, M., Kinnunen, P., Anderson, W., Ray, D. K., Puma, M. J., Varis, O., et al. (2023). Increased probability of hot and dry weather extremes during the growing season threatens global crop yields. *Sci. Rep.* 13, 3583. doi: 10.1038/s41598-023-29378-2

Heremans, S., Dong, Q., Zhang, B., Bydekerke, L., and Orshoven, J. V. (2015). Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and ¡i¿*in situ*¡/i¿ meteorological data. *J. Appl. Remote Sens.* 9, 97095. doi: 10.1117/1.JRS.9.097095

Hochberg, U., Rockwell, F. E., Holbrook, N. M., and Cochard, H. (2018). Iso/anisohydry: A plant–environment interaction rather than a simple hydraulic trait. *Trends Plant Sci.* 23, 112–120. doi: 10.1016/j.tplants.2017.11.002

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hoekstra, A. Y., and Mekonnen, M. M. (2012). The water footprint of humanity. *Proc. Natl. Acad. Sci.* 109, 3232–3237. doi: 10.1073/pnas.1109936109

Holman, F., Riche, A., Michalski, A., Castle, M., Wooster, M., and Hawkesford, M. (2016). High throughput field phenotyping of wheat plant height and growth rate in field plot trials using uav based remote sensing. *Remote Sens.* 8, 1031. doi: 10.3390/rs8121031

Holzinger, A., Keiblinger, K., Holub, P., Zatloukal, K., and Müller, H. (2023a). Ai for life: Trends in artificial intelligence for biotechnology. *New Biotechnol.* 74, 16–24. doi: 10.1016/j.nbt.2023.02.001

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdiscip. Reviews: Data Min. Knowledge Discovery* 9, e1312. doi: 10.1002/widm.1312

Holzinger, A., Malle, B., Saranti, A., and Pfeifer, B. (2021). Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. *Inf. Fusion* 71, 28–37. doi: 10.1016/j.inffus.2021.01.008

Holzinger, A., Saranti, A., Angerschmid, A., Finzel, B., Schmid, U., and Mueller, H. (2023b). Toward human-level concept learning: Pattern benchmarking for ai algorithms. *Patterns* 4, 100788. doi: 10.1016/j.patter.2023.100788

Holzinger, A., Saranti, A., Angerschmid, A., Retzlaff, C. O., Gronauer, A., Pejakovic, V., et al. (2022a). Digital transformation in smart farm and forest operations needs human-centered ai: Challenges and future directions. *Sensors* 22, 3043. doi: 10.3390/s22083043

Holzinger, A., Saranti, A., Hauschild, A.-C., Beinecke, J., Heider, D., Roettger, R., et al. (2023c). Humanin-the-loop integration with domain-knowledge graphs for explainable federated deep learning. *LNCS* 14065, 45–64. doi: 10.1007/978-3-031-40837-34

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022b). "Explainable ai methods - a brief overview," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (Vienna, Austria: Springer). Vol. 13200. 13–38. doi: 10.1007/978-3-031-04083-2_2/FIGURES/3

Holzworth, D. P., Huth, N. I., deVoil, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., et al. (2014). Apsim–evolution towards a new generation of agricultural systems simulation. *Environ. Model. Software* 62, 327–350. doi: 10.1016/j.envsoft.2014.07.009

Hu, Y., Knapp, S., and Schmidhalter, U. (2020). Advancing high-throughput phenotyping of wheat in early selection cycles. *Remote Sens.* 12, 574. doi: 10.3390/rs12030574

Huang, M., Liu, X., Zhou, Y., Summers, R. M., and Zhang, Z. (2019). Blink: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *GigaScience* 8, 1–12. doi: 10.1093/gigascience/giy154

Huang, S., Tang, L., Hupy, J. P., Wang, Y., and Shao, G. (2021). A commentary review on the use of normalized difference vegetation index (ndvi) in the era of popular remote sensing. *J. Forestry Res.* 32, 1–6. doi: 10.1007/s11676-020-01155-1

Huete, A. (1988). A soil-adjusted vegetation index (savi). *Remote Sens. Environ.* 25, 295–309. doi: 10.1016/0034-4257(88)90106-X

Hutchinson, M. L., Antono, E., Gibbons, B. M., Paradiso, S., Ling, J., and Meredig, B. (2017). doi: 10.48550/arXiv.1711.05099

Hütt, C., Bolten, A., Hüging, H., and Bareth, G. (2023). Uav lidar metrics for monitoring crop height, biomass and nitrogen uptake: A case study on a winter wheat field trial. *PFG - J. Photogrammetry Remote Sens. Geoinformation Sci.* 91, 65–76. doi: 10.1007/S41064-022-00228-6

Irwin, D. (2008). *The MassARRAY system for plant genomics* (Cambridge, Massachusetts, USA: CABI). doi: 10.1079/9781845933821.0098

Jayakodi, M., Schreiber, M., Stein, N., and Mascher, M. (2021). Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* 28, 1–9. doi: 10.1093/dnares/dsaa030

Jenal, A., Hüging, H., Ahrends, H. E., Bolten, A., Bongartz, J., and Bareth, G. (2021). Investigating the potential of a newly developed uav-mounted vnir/swir imaging system for monitoring crop traits—a case study for winter wheat. *Remote Sens.* 13, 1697. doi: 10.3390/rs13091697

Jeon, D., Kang, Y., Lee, S., Choi, S., Sung, Y., Lee, T.-H., et al. (2023). Digitalizing breeding in plants: A new trend of next-generation breeding based on genomic prediction. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1092584

Jiang, Y., and Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: A review. *Plant Phenomics* 2020, 4152816. doi: 10.34133/2020/4152816

John, M., Ankenbrand, M. J., Artmann, C., Freudenthal, J. A., Korte, A., and Grimm, D. G. (2022). Efficient permutation-based genome-wide association studies for normal and skewed phenotypic distributions. *Bioinformatics* 38, ii5–ii12. doi: 10.1093/bioinformatics/btac455

Joshi, A., Pradhan, B., Gite, S., and Chakraborty, S. (2023). Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: a systematic review. *Remote Sens.* 15, 2014. doi: 10.3390/rs15082014

Kadhim, N., Mourshed, M., and Bray, M. (2016). Advances in remote sensing applications for urban sustainability. *Euro-Mediterranean J. Environ. Integration* 1, 7. doi: 10.1007/s41207-016-0007-4

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., yee Kong, S., Freimer, N. B., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42, 348–354. doi: 10.1038/ng.548

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S. (2021). Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS J. Photogrammetry Remote Sens.* 173, 24–49. doi: 10.1016/j.isprsjprs.2020.12.010

Khadka, K., Earl, H. J., Raizada, M. N., and Navabi, A. (2020). A physio-morphological trait-based approach for breeding drought tolerant wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00715

Khan, Z., Rahimi-Eichi, V., Haefele, S., Garnett, T., and Miklavcic, S. J. (2018). Estimation of vegetation indices for high-throughput phenotyping of wheat using aerial imaging. *Plant Methods* 14, 20. doi: 10.1186/s13007-018-0287-6

Khanal, S., Fulton, J., and Shearer, S. (2017). An overview of current and potential applications of thermal remote sensing in precision agriculture. *Comput. Electron. Agric.* 139, 22–32. doi: 10.1016/j.compag.2017.05.001

Kim, J., Kim, S., Ju, C., and Son, H. I. (2019). Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications. *IEEE Access* 7, 105100–105115. doi: 10.1109/Access.6287639

Klepeckas, M., Januškaitienė, I., Vagusevičienė, I., and Juknys, R. (2020). Effects of different sowing time to phenology and yield of winter wheat. *Agric. Food Sci.* 29, 346–358. doi: 10.23986/afsci.90013

Koc, A., Odilbekov, F., Alamrani, M., Henriksson, T., and Chawade, A. (2022). Predicting yellow rust in wheat breeding trials by proximal phenotyping and machine learning. *Plant Methods* 18, 30. doi: 10.1186/s13007-022-00868-0

Koller, D., and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques* (Cambridge, Massachusetts, USA: MIT press).

Konvalina, P., Moudrý, J., Dotlačil, L., Stehno, Z., and Moudrý, J. (2010). Drought tolerance of land races of emmer wheat in comparison to soft wheat. *Cereal Res. Commun.* 38, 429–439. doi: 10.1556/CRC.38.2010.3.13

Koppensteiner, L. J., Kaul, H.-P., Piepho, H.-P., Barta, N., Euteneuer, P., Bernas, J., et al. (2022). Yield and yield components of facultative wheat are affected by sowing time, nitrogen fertilization and environment. *Eur. J. Agron.* 140, 126591. doi: 10.1016/j.eja.2022.126591

Korte, A., and Farlow, A. (2013). The advantages and limitations of trait analysis with gwas: a review. *Plant Methods* 9, 29. doi: 10.1186/1746-4811-9-29

Krasting, J. P., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., et al. (2018). NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP. *Earth System Grid Fed.* doi: 10.22033/ESGF/CMIP6.1407

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., et al. (2022). The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv [Preprint]*. arXiv:2202.01602. doi: 10.48550/arXiv.2202.01602

Kumar, D., Kushwaha, S., Delvento, C., Žilvinas, L., Vivekanand, V., Svensson, J. T., et al. (2020). Affordable phenotyping of winter wheat under field and controlled conditions for drought tolerance. *Agronomy* 10, 882. doi: 10.3390/agronomy10060882

Kumar, S., Röder, M. S., Singh, R. P., Kumar, S., Chand, R., Joshi, A. K., et al. (2016). Mapping of spot blotch disease resistance using ndvi as a substitute to visual observation in wheat (triticum aestivum l.). *Mol. Breed.* 36, 95. doi: 10.1007/s11032-016-0515-6

Lalic, B., Eitzinger, J., Mihailovic, D. T., Thaler, S., and Jancic, M. (2013). Climate change impacts on winter wheat yield change – which climatic parameters are crucial in pannonian lowland? *J. Agric. Sci.* 151, 757–774. doi: 10.1017/S0021859612000640

Langridge, P., and Reynolds, M. (2021). Breeding for drought and heat tolerance in wheat. *Theor. Appl. Genet.* 134, 1753–1769. doi: 10.1007/s00122-021-03795-1

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* 10, 1096. doi: 10.1038/s41467-019-08987-4

Lehnert, H., Berner, T., Lang, D., Beier, S., Stein, N., Himmelbach, A., et al. (2022). Insights into breeding history, hotspot regions of selection, and untapped allelic diversity for bread wheat breeding. *Plant J.* 112, 897–918. doi: 10.1111/tpj.15952

Lehnert, H., Serfling, A., Friedt, W., and Ordon, F. (2018). Genome-wide association studies reveal genomic regions associated with the response of wheat (triticum aestivum l.) to mycorrhizae under drought stress conditions. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01728

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat* 9, 1350–1371. doi: 10.1214/15-AOAS848

Levitt, J. (1980). *Responses of plants to environmental stresses. Volume II. Water, radiation, salt, and other stresses*. 1, 497.

Li, A., Hao, C., Wang, Z., Geng, S., Jia, M., Wang, F., et al. (2022). Wheat breeding history reveals synergistic selection of pleiotropic genomic sites for plant architecture and grain yield. *Mol. Plant* 15, 504–519. doi: 10.1016/j.molp.2022.01.004

Li, C., Li, L., Reynolds, M. P., Wang, J., Chang, X., Mao, X., et al. (2021). Recognizing the hidden half in wheat: root system attributes associated with drought tolerance. *J. Exp. Bot.* 72, 5117–5133. doi: 10.1093/jxb/erab124

Li, H., Zhang, Y., Lei, Y., Antoniuk, V., and Hu, C. (2019). Evaluating different non-destructive estimation methods for winter wheat (triticum aestivum l.) nitrogen status based on canopy spectrum. *Remote Sens.* 12, 95. doi: 10.3390/rs12010095

Lichthardt, C., Chen, T.-W., Stahl, A., and Stützel, H. (2020). Co-evolution of sink and source in the recent breeding history of winter wheat in Germany. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01771

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., et al. (2012). Gapit: genome association and prediction integrated tool. *Bioinformatics* 28, 2397–2399. doi: 10.1093/bioinformatics/bts444

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PloS Genet.* 12, e1005767. doi: 10.1371/journal.pgen.1005767

Liu, G., Liu, D., Zhang, A., Liu, H., Mia, M. S., Mullan, D., et al. (2023). Identification of kasp markers and candidate genes for drought tolerance in wheat using 90k snp array genotyping of nearisogenic lines targeting a 4bs quantitative trait locus. *Theor. Appl. Genet.* 136, 1–13. doi: 10.1007/s00122-023-04438-3

Lopes, M. S., and Reynolds, M. P. (2010). Partitioning of assimilates to deeper roots is associated with cooler canopies and increased yield under drought in wheat. *Funct. Plant Biol.* 37, 147. doi: 10.1071/FP09121

López-Cortegano, E., and Caballero, A. (2019). Inferring the nature of missing heritability in human traits using data from the gwas catalog. *Genetics* 212, 891–904. doi: 10.1534/genetics.119.302077

Lu, Y., Chen, D., Olaniyi, E. O., and Huang, Y. (2022). Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Comput. Electron. Agric.* 200, 107208. doi: 10.1016/j.compag.2022.107208

Lunzer, M., Buerstmayr, M., Grausgruber, H., Müllner, A. E., Fallbacher, I., and Buerstmayr, H. (2023). Wheat (triticum aestivum) chromosome 6d harbours the broad spectrum common bunt resistance gene bt11. *Theor. Appl. Genet.* 136, 1–17. doi: 10.1007/s00122-023-04452-5

Lüttger, A. B., and Feike, T. (2018). Development of heat and drought related extreme weather events and their effect on winter wheat yields in Germany. *Theor. Appl. Climatology* 132, 15–29. doi: 10.1007/s00704-017-2076-y

Ma, J., Zhao, D., Tang, X., Yuan, M., Zhang, D., Xu, M., et al. (2022). Genome-wide association study on root system architecture and identification of candidate genes in wheat (triticum aestivum l.). *Int. J. Mol. Sci.* 23, 1843. doi: 10.3390/ijms23031843

Mackay, I., and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 12, 57–63. doi: 10.1016/j.tplants.2006.12.001

Magister, L. C., Kazhdan, D., Singh, V., and Liò, P. (2021). Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *arXiv [Preprint]*. arXiv:2107.11889. doi: 10.48550/arXiv.2107.11889

Mahlein, A.-K., Oerke, E.-C., Steiner, U., and Dehne, H.-W. (2012). Recent advances in sensing plant diseases for precision crop protection. *Eur. J. Plant Pathol.* 133, 197–209. doi: 10.1007/s10658-011-9878-z

Makhoul, M., Rambla, C., Voss-Fels, K. P., Hickey, L. T., Snowdon, R. J., and Obermeier, C. (2020). Overcoming polyploidy pitfalls: a user guide for effective snp conversion into kasp markers in wheat. *Theor. Appl. Genet.* 133, 2413–2430. doi: 10.1007/s00122-020-03608-x

Manickavelu, A., Hattori, T., Yamaoka, S., Yoshimura, K., Kondou, Y., Onogi, A., et al. (2017). Genetic nature of elemental contents in wheat grains and its genomic prediction: Toward the effective use of wheat landraces from Afghanistan. *PloS One* 12, e0169416. doi: 10.1371/journal.pone.0169416

MansChadi, A. M., Hammer, G. L., Christopher, J. T., and deVoil, P. (2008). Genotypic variation in seedling root architectural traits and implications for drought adaptation in wheat (triticum aestivum l.). *Plant Soil* 303, 115–129. doi: 10.1007/s11104-007-9492-1

March, R. E. (1999). Gene mapping by linkage and association analysis. *Appl. Biochem. Biotechnol. - Part B Mol. Biotechnol.* 13, 113–122. doi: 10.1385/MB:13:2:113/METRICS

Marchini, J. (2010). SNPTEST v2 Technical Details *()*.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* 39, 906–913. doi: 10.1038/ng2088

Marian, A. J. (2012). Elements of 'missing heritability'. *Curr. Opin. Cardiol.* 27, 197–201. doi: 10.1097/HCO.0b013e328352707d

Maseda, P. H., and Fernandez, R. J. (2006). Stay wet or else: three ways in which plants can adjust hydraulically to their environment. *J. Exp. Bot.* 57, 3963–3977. doi: 10.1093/jxb/erl127

Matese, A., Czarnecki, J. M. P., Samiappan, S., and Moorhead, R. (2023). Are unmanned aerial vehiclebased hyperspectral imaging and machine learning advancing crop science? *Trends Plant Sci.* 29, 196–209. doi: 10.1016/j.tplants.2023.09.001

Mathew, I., Shimelis, H., Shayanowako, A. I. T., Laing, M., and Chaplot, V. (2019). Genome-wide association study of drought tolerance and biomass allocation in wheat. *PloS One* 14, e0225383. doi: 10.1371/journal.pone.0225383

Matsushita, B., Yang, W., Chen, J., Onda, Y., and Qiu, G. (2007). Sensitivity of the enhanced vegetation index (evi) and normalized difference vegetation index (ndvi) to topographic effects: A case study in high-density cypress forest. *Sensors* 7, 2636–2651. doi: 10.3390/s7112636

Maulana, F., Huang, W., Anderson, J. D., and Ma, X. F. (2020). Genome-wide association mapping of seedling drought tolerance in winter wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.573786

Miao, C., Yang, J., and Schnable, J. C. (2019). Optimising the identification of causal variants across varying genetic architectures in crops. *Plant Biotechnol. J.* 17, 893–905. doi: 10.1111/pbi.13023

Miedaner, T., Wilde, F., Steiner, B., Buerstmayr, H., Korzun, V., and Ebmeyer, E. (2006). Stacking quantitative trait loci (qtl) for fusarium head blight resistance from non-adapted sources in an european elite spring wheat background and assessing their effects on deoxynivalenol (don) content and disease severity. *Theor. Appl. Genet.* 112, 562–569. doi: 10.1007/s00122-005-0163-4

Mieth, B., Kloft, M., Rodr´ıguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genomewide association studies. *Sci. Rep.* 6, 36671. doi: 10.1038/SREP36671

Mieth, B., Rozier, A., Rodriguez, J. A., Höhne, M. M. C., Görnitz, N., and Müller, K.-R. (2021). Deepcombi: explainable artificial intelligence for the analysis and discovery in genome-wide association studies. *NAR Genomics Bioinf.* 3, Iqab065. doi: 10.1093/nargab/lqab065

Miller, T., Hoffman, R., Amir, O., and Holzinger, A. (2022). Special issue on explainable artificial intelligence (xai). *Artif. Intell.* 307, 103705. doi: 10.1016/j.artint.2022.103705

Mills, M. C., and Rahal, C. (2019). A scientometric review of genome-wide association studies. *Commun. Biol.* 2, 9. doi: 10.1038/s42003-018-0261-x

Mohammadi, M., Xavier, A., Beckett, T., Beyer, S., Chen, L., Chikssa, H., et al. (2020). Identification, deployment, and transferability of quantitative trait loci from genome-wide association studies in plants. *Curr. Plant Biol.* 24, 100145. doi: 10.1016/j.cpb.2020.100145

Mohanty, S. P., Hughes, D. P., and Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01419

Mohsan, S. A. H., Othman, N. Q. H., Li, Y., Alsharif, M. H., and Khan, M. A. (2023). Unmanned aerial vehicles (uavs): practical aspects, applications, open challenges, security issues, and future trends. *Intelligent Service Robotics*. 16, 109-137. doi: 10.1007/s11370-022-00452-4

Molero, G., Joynson, R., Pinera-Chavez, F. J., Gardiner, L., Rivera-Amado, C., Hall, A., et al. (2019). Elucidating the genetic basis of biomass accumulation and radiation use efficiency in spring wheat and its role in yield potential. *Plant Biotechnol. J.* 17, 1276–1288. doi: 10.1111/pbi.13052

Molin, E. M. (2024). "Wheatviz - an interdisciplinary project to accelerate the breeding of drought-tolerant winter wheat via cutting-edge genomics and uav-based phenotyping by integrating explainable ai," in *Proceedings of the 74th Conference of the Vereinigung der Pflanzenzüchter und Saatgutkaufleute Österreichs*, Raumberg-Gumpenstein, Irdning, Austria, 20-22 November 2023 (Vienna, Austria: University of Natural Resources and Life Sciences), 11-12.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. (2019). "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer International Publishing, Cham), 193–209. doi: 10.1007/978-3-030-28954-610

Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C. K., et al. (2017). The pangenome of hexaploid bread wheat. *Plant J.* 90, 1007–1013. doi: 10.1111/tpj.13515

Mori, M., Inagaki, M., Inoue, T., and Nachit, M. (2011). Association of root water-uptake ability with drought adaptation in wheat. *Cereal Res. Commun.* 39, 551–559. doi: 10.1556/CRC.39.2011.4.10

Muhammad, M. (2021). Artificial Intelligence Algorithms for Polygenic Genotype-Phenotype Predictions. [Master's thesis]. The University of Queensland.

Mukarram, M., Choudhary, S., Kurjak, D., Petek, A., and Khan, M. M. A. (2021). Drought: Sensing, signalling, effects and tolerance in higher plants. *Physiologia Plantarum* 172, 1291–1300. doi: 10.1111/ppl.13423

Mullan, D. J., and Reynolds, M. P. (2010). Quantifying genetic effects of ground cover on soil water evaporation using digital imaging. *Funct. Plant Biol.* 37, 703. doi: 10.1071/FP09277

Müller, H., Holzinger, A., Plass, M., Brcic, L., Stumptner, C., and Zatloukal, K. (2022). Explainability and causability for artificial intelligence-supported medical image analysis in the context of the european *in vitro* diagnostic regulation. *New Biotechnol.* 70, 67–72. doi: 10.1016/j.nbt.2022.05.002

Mwadzingeni, L., Shimelis, H., Dube, E., Laing, M. D., and Tsilo, T. J. (2016). Breeding wheat for drought tolerance: Progress and technologies. *J. Integr. Agric.* 15, 935–943. doi: 10.1016/S2095-3119(15)61102-9

Mwadzingeni, L., Shimelis, H., Rees, D. J. G., and Tsilo, T. J. (2017). Genome-wide association analysis of agronomic traits in wheat under drought-stressed and non-stressed conditions. *PloS One* 12, e0171692. doi: 10.1371/journal.pone.0171692

Myneni, R. B., Hall, F. G., Sellers, P. J., and Marshak, A. L. (1995). The interpretation of spectral vegetation indexes. *IEEE Trans. Geosci. Remote Sens.* 33, 481–486. doi: 10.1109/TGRS.1995.8746029

Najafabadi, M. Y., Hesami, M., and Eskandari, M. (2023). Machine learning-assisted approaches in modernized plant breeding programs. *Genes* 14, 777. doi: 10.3390/genes14040777

Nakhforoosh, A., Grausgruber, H., Kaul, H.-P., and Bodner, G. (2014). Wheat root diversity and root functional characterization. *Plant Soil* 380, 211–229. doi: 10.1007/s11104-014-2082-0

Nakhforoosh, A., Grausgruber, H., Kaul, H.-P., and Bodner, G. (2015). Dissection of drought response of modern and underutilized wheat varieties according to passioura's yield-water framework. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00570

Napierala, M. A. (2012). What is the bonferroni correction? *AAOS Now*, 40–41.

NASA (2023) *Landsat science*. Available online at: https://landsat.gsfc.nasa.gov/.

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Stat. Society. Ser. A (General)* 135, 370. doi: 10.2307/2344614

Neugschwandtner, R. W., Böhm, K., Hall, R. M., and Kaul, H. P. (2015). Development, growth, and nitrogen use of autumn- and spring-sown facultative wheat. *Acta Agriculturae Scandinavica Section B — Soil Plant Sci.* 65, 6–13. doi: 10.1080/09064710.2014.958522

Nguyen, C., Sagan, V., Skobalski, J., and Severo, J. I. (2023). Early detection of wheat yellow rust disease and its impact on terminal yield with multi-spectral uav-imagery. *Remote Sens.* 15, 3301. doi: 10.3390/rs15133301

Niazian, M., and Niedbała, G. (2020). Machine learning for plant breeding and biotechnology. *Agriculture* 10, 436. doi: 10.3390/agriculture10100436

Nicholls, H. L., John, C. R., Watson, D. S., Munroe, P. B., Barnes, M. R., and Cabrera, C. P. (2020). Reaching the end-game for gwas: Machine learning approaches for the prioritization of complex disease loci. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00350

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Front. Genet.* 6. doi: 10.3389/fgene.2015.00285

Nilson, T. (1971). A theoretical analysis of the frequency of gaps in plant stands. *Agric. Meteorology* 8, 25–38. doi: 10.1016/0002-1571(71)90092-6

Nordborg, M., and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18, 83–90. doi: 10.1016/S0168-9525(02)02557-X

Oliveira, F. A., Jones, J. W., Pavan, W., Bhakta, M., Vallejos, C. E., Correll, M. J., et al. (2021). Incorporating a dynamic gene-based process module into a crop simulation model. *silico Plants* 3, diab011. doi: 10.1093/insilicoplants/diab011

Ozturk, A., Caglar, O., and Bulut, S. (2006). Growth and yield response of facultative wheat to winter sowing, freezing sowing and spring sowing at different seeding rates. *J. Agron. Crop Sci.* 192, 10–16. doi: 10.1111/j.1439-037X.2006.00187.x

Palta, J. A., Chen, X., Milroy, S. P., Rebetzke, G. J., Dreccer, M. F., and Watt, M. (2011). Large root systems: are they useful in adapting wheat to dry environments? *Funct. Plant Biol.* 38, 347. doi: 10.1071/FP11031

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-53451-4

Passioura, J. B. (1977). Grain yield, harvest index, and water use of wheat. *J. Aust. Institute Agric. Sci.* 43, 117–120.

Patidar, A., Yadav, M. C., Kumari, J., Tiwari, S., Chawla, G., and Paul, V. (2023). Identification of climate-smart bread wheat germplasm lines with enhanced adaptation to global warming. *Plants* 12, 2851. doi: 10.3390/plants12152851

Pearl, J., and Mackenzie, D. (2018). The book of why: The new science of cause and effect. *Science* 361, 855–855. doi: 10.1126/science.aau9731

Peng, Z. (2020). A brief overview of gwas: discover genetic variations of diseases and phenotypes. *E3S Web Conferences* 185, 3014. doi: 10.1051/e3sconf/202018503014

Pfeifer, B., Holzinger, A., and Schimek, M. G. (2022). Robust random forest-based all-relevant feature ranks for trustworthy ai. *Stud. Health Technol. Inf.* 294, 137–138. doi: 10.3233/SHTI220418

Pierce, S. E., Booms, A., Prahl, J., van der Schans, E. J. C., Tyson, T., and Coetzee, G. A. (2020). Post-gwas knowledge gap: the how, where, and when. *NPJ Parkinson's Dis.* 6, 23. doi: 10.1038/s41531-020-00125-y

Pieruschka, R., and Schurr, U. (2019). Plant phenotyping: Past, present, and future. *Plant Phenomics* 2019, 7507131. doi: 10.34133/2019/7507131

Prabhakar, M., Prasad, Y. G., and Rao, M. N. (2012). "Remote sensing of biotic stress in crop plants and its applications for pest management," in *Crop Stress and its Management: Perspectives and Strategies* (Springer Netherlands, Dordrecht), 517–545. doi: 10.1007/978-94-007-2220-016

Prudnikova, E., Savin, I., Vindeker, G., Grubina, P., Shishkonakova, E., and Sharychev, D. (2019). Influence of soil background on spectral reflectance of winter wheat crop canopy. *Remote Sens.* 11, 1932. doi: 10.3390/rs11161932

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). Plink: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Qaseem, M. F., Qureshi, R., Muqaddasi, Q. H., Shaheen, H., Kousar, R., and Röder, M. S. (2018). Genomewide association mapping in bread wheat subjected to independent and combined high temperature and drought stress. *PloS One* 13, e0199121. doi: 10.1371/journal.pone.0199121

Qaseem, M. F., Qureshi, R., Shaheen, H., and Shafqat, N. (2019). Genome-wide association analyses for yield and yield-related traits in bread wheat (triticum aestivum l.) under pre-anthesis combined heat and drought stress in field conditions. *PloS One* 14, e0213407. doi: 10.1371/journal.pone.0213407

Qi, J., Chehbouni, A., Huete, A., Kerr, Y., and Sorooshian, S. (1994). A modified soil adjusted vegetation index. *Remote Sens. Environ.* 48, 119–126. doi: 10.1016/0034-4257(94)90134-1

Qin, T., Feng, J., Zhang, X., Li, C., Fan, J., Zhang, C., et al. (2023). Continued decline of global soil moisture content, with obvious soil stratification and regional difference. *Sci. Total Environ.* 864, 160982. doi: 10.1016/j.scitotenv.2022.160982

Quarrie, S. A., Steed, A., Calestani, C., Semikhodskii, A., Lebreton, C., Chinoy, C., et al. (2005). A high-density genetic map of hexaploid wheat (triticum aestivum l.) from the cross chinese spring × sq1 and its use to compare qtls for grain yield across a range of environments. *Theor. Appl. Genet.* 110, 865–880. doi: 10.1007/s00122-004-1902-7

Radočaj, D., Šiljeg, A., Marinovič, R., and Jurišić, M. (2023). State of major vegetation indices in precision agriculture studies indexed in web of science: A review. *Agriculture* 13, 707. doi: 10.3390/agriculture13030707

Rahimi, Y., Khahani, B., Jamali, A., Alipour, H., Bihamta, M. R., and Ingvarsson, P. K. (2023). Genomewide association study to identify genomic loci associated with early vigor in bread wheat under simulated water deficit complemented with quantitative trait loci meta-analysis. *G3 Genes—Genomes—Genetics* 13, jkac320. doi: 10.1093/g3journal/jkac320

Rahman, M. M., and Robson, A. J. (2016). A novel approach for sugarcane yield prediction using landsat time series imagery: A case study on bundaberg region. *Adv. Remote Sens.* 05, 93–102. doi: 10.4236/ars.2016.52008

Ramstein, G. P., Jensen, S. E., and Buckler, E. S. (2019). Breaking the curse of dimensionality to identify causal variants in breeding 4. *Theor. Appl. Genet.* 132, 559–567. doi: 10.1007/s00122-018-3267-3

Rasheed, A., Wen, W., Gao, F., Zhai, S., Jin, H., Liu, J., et al. (2016). Development and validation of kasp assays for genes underpinning key economic traits in bread wheat. *Theor. Appl. Genet.* 129, 1843–1860. doi: 10.1007/s00122-016-2743-x

Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D. N., et al. (2020). Breeder friendly phenotyping. *Plant Sci.* 295, 110396. doi: 10.1016/j.plantsci.2019.110396

Reynolds, M., Dreccer, F., and Trethowan, R. (2006). Drought-adaptive traits derived from wheat wild relatives and landraces. *J. Exp. Bot.* 58, 177–186. doi: 10.1093/jxb/erl250

Reynolds, M., and Langridge, P. (2016). Physiological breeding. *Curr. Opin. Plant Biol.* 31, 162–171. doi: 10.1016/j.pbi.2016.04.005

Rezaei, E. E., Siebert, S., Manderscheid, R., Müller, J., Mahrookashani, A., Ehrenpfordt, B., et al. (2018). Quantifying the response of wheat yields to heat stress: The role of the experimental setup. *Field Crops Res.* 217, 93–103. doi: 10.1016/j.fcr.2017.12.015

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., et al. (2017). The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environ. Change* 42, 153–168. doi: 10.1016/j.gloenvcha.2016.05.009

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-August-2016. (New York, NY, United States), 1135–1144. doi: 10.1145/2939672.2939778

Richards, R. A. (2006). Physiological traits used in the breeding of new cultivars for water-scarce environments. *Agric. Water Manage.* 80, 197–211. doi: 10.1016/j.agwat.2005.07.013

Rimbert, H., Darrier, B., Navarro, J., Kitt, J., Choulet, F., Leveugle, M., et al. (2018). High throughput snp discovery and genotyping in hexaploid wheat. *PloS One* 13, e0186329. doi: 10.1371/journal.pone.0186329

Romero, J. C. (2022). Better understanding genomic architecture with the use of applied statistics and explainable artificial intelligence. [Phd thesis] University of Tennessee.

Rouse, J. W., Haas, R. H., Schell, J. A., and Deering, D. W. (1974). Monitoring vegetation systems in the great plains with erts. *NASA Spec. Publ* 351, 309.

Sabatti, C. (2013). *Multivariate linear models for gwas* (Cambridge, UK: Cambridge University Press), 188–207. doi: 10.1017/CBO9781139226448.010

Safavian, S., and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Trans. Systems Man Cybernetics* 21, 660–674. doi: 10.1109/21.97458

Saini, D. K., Chopra, Y., Singh, J., Sandhu, K. S., Kumar, A., Bazzer, S., et al. (2022). Comprehensive evaluation of mapping complex traits in wheat using genome-wide association studies. *Mol. Breed.* 42, 1. doi: 10.1007/s11032-021-01272-7

Sallam, A., Alqudah, A. M., Dawood, M. F. A., Baenziger, P. S., and Börner, A. (2019). Drought stress tolerance in wheat and barley: Advances in physiology, breeding and genetics research. *Int. J. Mol. Sci.* 20, 3137. doi: 10.3390/ijms20133137

Sanguineti, M., Li, S., Maccaferri, M., Corneti, S., Rotondo, F., Chiari, T., et al. (2007). Genetic dissection of seminal root architecture in elite durum wheat germplasm. *Ann. Appl. Biol.* 151, 291–305. doi: 10.1111/j.1744-7348.2007.00198.x

Santorsola, M., and Lescai, F. (2023). The promise of explainable deep learning for omics data analysis: Adding new discovery tools to ai. *New Biotechnol.* 77, 1–11. doi: 10.1016/j.nbt.2023.06.002

Saranti, A., Hudec, M., Mináriková, E., Takáč, Z., Großschedl, U., Koch, C., et al. (2022). Actionable explainable ai (axai): A practical example with aggregation functions for adaptive classification and textual explanations for interpretable machine learning. *Mach. Learn. Knowledge Extraction* 4, 924–953. doi: 10.3390/make4040047

Saranya, T., Deisy, C., Sridevi, S., and Anbananthen, K. S. M. (2023). A comparative study of deep learning and internet of things for precision agriculture. *Eng. Appl. Artif. Intell.* 122, 106034. doi: 10.1016/j.engappai.2023.106034

SAS Institute Inc (2013). *JMP genomics, Version 6.0*. Cary, North Carolina: SAS Institute Inc.

Satorre, E. H., and Slafer, G. A. (2000). *Wheat: ecology and physiology of yield determination* (Binghamton, NY, USA: Food Products Press).

Schils, R., Olesen, J. E., Kersebaum, K.-C., Rijk, B., Oberforster, M., Kalyada, V., et al. (2018). Cereal yield gaps across europe. *Eur. J. Agron.* 101, 109–120. doi: 10.1016/j.eja.2018.09.003

Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., et al. (2020). Higher-order explanations of graph neural networks via relevant walks. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 7581–7596. doi: 10.1109/TPAMI.2021.3115452

Schwalbe, G., and Finzel, B. (2023). "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," in *Data Mining and Knowledge Discovery*. (Cham, Switzerland: Springer). doi: 10.1007/s10618-022-00867-8

Semagn, K., Babu, R., Hearne, S., and Olsen, M. (2014). Single nucleotide polymorphism genotyping using kompetitive allele specific pcr (kasp): overview of the technology and its application in crop improvement. *Mol. Breed.* 33, 1–14. doi: 10.1007/s11032-013-9917-x

Senapati, N., Halford, N. G., and Semenov, M. A. (2021). Vulnerability of european wheat to extreme heat and drought around flowering under future climate. *Environ. Res. Lett.* 16, 024052. doi: 10.1088/1748-9326/abdcf3

Shaikh, T. A., Rasool, T., and Lone, F. R. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. *Comput. Electron. Agric.* 198, 107119. doi: 10.1016/j.compag.2022.107119

Shapley, L. S. (1952). *A value for n-person games*. Princeton, New Jersey: Princeton University Press. doi: 10.7249/P0295

Sharma, D., Durand, A., Legault, M.-A., Perreault, L.-P. L., Lemac¸on, A., Dubé, M.-P., et al. (2020). Deep interpretability for gwas. *arXiv [Preprint]*. arXiv:2007.01516. doi: 10.48550/arXiv.2007.01516

Shi, X., and Ling, H.-Q. (2018). Current advances in genome sequencing of common wheat and its ancestral species. *Crop J.* 6, 15–21. doi: 10.1016/j.cj.2017.11.001

Shokat, S., Großkinsky, D. K., Singh, S., and Liu, F. (2023). The role of genetic diversity and pre-breeding traits to improve drought and heat tolerance of bread wheat at the reproductive stage. *Food Energy Secur.* 12, e478. doi: 10.1002/fes3.478

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi: 10.1186/s40537-019-0197-0

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning* (Sydney, Australia: PMLR), Vol. 70, 4844–4866.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*.

Sishodia, R. P., Ray, R. L., and Singh, S. K. (2020). Applications of remote sensing in precision agriculture: A review. *Remote Sens.* 12, 3136. doi: 10.3390/rs12193136

Sokol, K., and Flach, P. (2019). Desiderata for interpretability: Explaining decision tree predictions with counterfactuals. *Proc. AAAI Conf. Artif. Intell.* 33, 10035–10036. doi: 10.1609/aaai.v33i01.330110035

Song, P., Wang, J., Guo, X., Yang, W., and Zhao, C. (2021). High-throughput phenotyping: Breaking through the bottleneck in future crop breeding. *Crop J.* 9, 633–645. doi: 10.1016/j.cj.2021.03.015

Song, L., Wang, R., Yang, X., Zhang, A., and Liu, D. (2023). Molecular markers and their applications in marker-assisted selection (mas) in bread wheat (triticum aestivum l.). *Agriculture* 13, 642. doi: 10.3390/agriculture13030642

Sperry, J. S., Adler, F. R., Campbell, G. S., and Comstock, J. P. (1998). Limitation of plant water use by rhizosphere and xylem conductance: results from a model. *Plant Cell Environ.* 21, 347–359. doi: 10.1046/j.1365-3040.1998.00287.x

Sperry, J. S., and Love, D. M. (2015). What plant hydraulics can tell us about responses to climate-change droughts. *New Phytol.* 207, 14–27. doi: 10.1111/nph.13354

Spiertz, J. (1974). Grain growth and distribution of dry matter in the wheat plant as influenced by temperature, light energy and ear size. *Netherlands J. Agric. Sci.* 22, 207–220. doi: 10.18174/njas.v22i3.17223

Sposito, G. (2013). Green water and global food security. *Vadose Zone J.* 12, 1–6. doi: 10.2136/vzj2013.02.0041

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., et al. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Sci. Rep.* 12, 1–14. doi: 10.1038/s41598-022-06249-w

Staniak, M., and Biecek, P. (2019). Explanations of model predictions with live and breakdown packages. *R J.* 10, 395. doi: 10.32614/RJ-2018-072

Stein, B. V., Raponi, E., Sadeghi, Z., Bouman, N., Ham, R. C. V., and Back, T. (2022). A comparison of global sensitivity analysis methods for explainable ai with an application in genomic prediction. *IEEE Access* 10, 103364–103381. doi: 10.1109/ACCESS.2022.3210175

Steiner, B., Buerstmayr, M., Michel, S., Schweiger, W., Lemmens, M., and Buerstmayr, H. (2017). Breeding strategies and advances in line selection for fusarium head blight resistance in wheat. *Trop. Plant Pathol.* 42, 165–174. doi: 10.1007/s40858-017-0127-7

Stephens, M., and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10, 681–690. doi: 10.1038/nrg2615

Streich, J., Romero, J., Gazolla, J. G. F. M., Kainer, D., Cliff, A., Prates, E. T., et al. (2020). Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the united nations sustainable development goals? *Curr. Opin. Biotechnol.* 61, 217–225. doi: 10.1016/j.copbio.2020.01.010

Stutsel, B., Johansen, K., Malbéteau, Y. M., and McCabe, M. F. (2021). Detecting plant stress using thermal and optical imagery from an unoccupied aerial vehicle. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.734944

Subedi, M., Ghimire, B., Bagwell, J. W., Buck, J. W., and Mergoum, M. (2023). Wheat end-use quality: State of art, genetics, genomics-assisted improvement, future challenges, and opportunities. *Front. Genet.* 13. doi: 10.3389/fgene.2022.1032601

Suliman, S., Alemu, A., Abdelmula, A. A., Badawi, G. H., Al-Abdallat, A., and Tadesse, W. (2021). Genome-wide association analysis uncovers stable qtls for yield and quality traits of spring bread wheat (triticum aestivum) across contrasting environments. *Plant Gene* 25, 100269. doi: 10.1016/j.plgene.2020.100269

Sun, D., Cen, H., Weng, H., Wan, L., Abdalla, A., El-Manawy, A. I., et al. (2019). Using hyperspectral analysis as a potential high throughput phenotyping tool in gwas for protein content of rice quality. *Plant Methods* 15, 54. doi: 10.1186/s13007-019-0432-x

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N., and Chen, F. (2020). The wheat 660k snp array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant Biotechnol. J.* 18, 1354–1360. doi: 10.1111/pbi.13361

Sun, H., Wang, B., Wu, Y., and Yang, H. (2023). Deep learning method based on spectral characteristic rein-forcement for the extraction of winter wheat planting area in complex agricultural landscapes. *Remote Sens.* 15, 1301. doi: 10.3390/rs15051301

Syvänen, A.-C. (2005). Toward genome-wide snp genotyping. *Nat. Genet.* 37, S5–S10. doi: 10.1038/ng1558

Szymczak, S., Biernacka, J. M., Cordell, H. J., González-Recio, O., König, I. R., Zhang, H., et al. (2009). Machine learning in genome-wide association studies. *Genet. Epidemiol.* 33, 51-57. doi: 10.1002/gepi.20473

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* 20, 467–484. doi: 10.1038/s41576-019-0127-1

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). Gapit version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome* 9. doi: 10.3835/plantgenome2015.11.0120

Tang, Z., Sun, Y., Wan, G., Zhang, K., Shi, H., Zhao, Y., et al. (2022). Winter wheat lodging area extraction using deep learning with gaofen-2 satellite imagery. *Remote Sens.* 14, 4887. doi: 10.3390/rs14194887

Tang, Z., Wang, M., Schirrmann, M., Dammer, K.-H., Li, X., Brueggeman, R., et al. (2023). Affordable high throughput field detection of wheat stripe rust using deep learning with semi-automated image labeling. *Comput. Electron. Agric.* 207, 107709. doi: 10.1016/j.compag.2023.107709

Tardieu, F., and Simonneau, T. (1998). Variability among species of stomatal control under fluctuating soil water status and evaporative demand: modelling isohydric and anisohydric behaviours. *J. Exp. Bot.* 49, 419–432. doi: 10.1093/jxb/49.Special_Issue.419

Tattaris, M., Reynolds, M. P., and Chapman, S. C. (2016). A direct comparison of remote sensing approaches for high-throughput phenotyping in plant breeding. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01131

Technow, F., Messina, C. D., Totir, L. R., and Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate bayesian computation. *PloS One* 10, e0130855. doi: 10.1371/journal.pone.0130855

Temenos, A., Tzortzis, I. N., Kaselimi, M., Rallis, I., Doulamis, A., and Doulamis, N. (2022). Novel insights in spatial epidemiology utilizing explainable ai (xai) and remote sensing. *Remote Sens.* 14, 3074. doi: 10.3390/rs14133074

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of dna sequence polymorphism along chromosome 1 of maize (¡i¿zea mays¡/i¿ ssp. ¡i¿mays¡/i¿ l.). *Proc. Natl. Acad. Sci.* 98, 9161–9166. doi: 10.1073/pnas.151244298

Thenkabail, P. S., Lyon, J. G., and Huete, A. (2018). *Biophysical and biochemical characterization and plant species studies* (Boca Raton, Florida, USA: CRC Press). doi: 10.1201/9780429431180

Thenkabail, P. S., Smith, R. B., and De Pauw, E. (2002). Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization. *Photogrammetric Eng. Remote Sens.* 68, 607–622.

Thoday-Kennedy, E., Good, N., and Kant, S. (2022). "Basics of sensor-based phenotyping in wheat," in *Accelerated Breeding of Cereal Crops* (Springer US, New York, NY), 305–331. doi: 10.1007/978-1-0716-1526-316

Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* 28, 286–289. doi: 10.1038/90135

Tilling, A. K, O'Leary, G. J., Ferwerda, J. G., Jones, S. D., Fitzgerald, G. J., Rodriguez, D., et al. (2007). Remote sensing of nitrogen and water stress in wheat. *Field Crops Res.* 104, 77–85. doi: 10.1016/j.fcr.2007.03.023

Trethowan, R. M., and Mujeeb-Kazi, A. (2008). Novel germplasm resources for improving environmental stress tolerance of hexaploid wheat. *Crop Sci.* 48, 1255–1265. doi: 10.2135/cropsci2007.08.0477

Trnka, M., Olesen, J. E., Kersebaum, K. C., Skjelvåg, A. O., Eitzinger, J., Seguin, B., et al. (2011). Agroclimatic conditions in europe under climate change. *Global Change Biol.* 17, 2298–2318. doi: 10.1111/j.1365-2486.2011.02396.x

Turner, N. C., Blum, A., Cakir, M., Steduto, P., Tuberosa, R., and Young, N. (2014). Strategies to increase the yield and yield stability of crops under drought – are we making progress? *Funct. Plant Biol.* 41, 1199. doi: 10.1071/FP14057

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., et al. (2021). Genomewide association studies. *Nat. Rev. Methods Primers* 1, 59. doi: 10.1038/s43586-021-00056-9

Unesco (2012). *Managing Water Under Uncertainty and Risk* Vol. 1 (Paris, France: Unesco).

Vadez, V. (2014). Root hydraulics: The forgotten side of roots in drought adaptation. *Field Crops Res.* 165, 15–24. doi: 10.1016/j.fcr.2014.03.017

Vadez, V., Kholova, J., Zaman-Allah, M., and Belko, N. (2013). Water: the most important 'molecular' component of water stress tolerance research. *Funct. Plant Biol.* 40, 1310. doi: 10.1071/FP13149

Vadez, V., Krishnamurthy, L., Kashiwagi, J., Kholova, J., Devi, J. M., Sharma, K. K., et al. (2007). Exploiting the functionality of root systems for dry, saline, and nutrient deficient environments in a changing climate. *J. SAT Agric. Res.* 4, 1–61.

Valkoun, J. J. (2001). Wheat pre-breeding using wild progenitors. *Euphytica* 119, 17–23. doi: 10.1023/A:1017562909881

van der Velde, M., Baruth, B., Bussay, A., Ceglar, A., Condado, S. G., Karetsos, S., et al. (2018). In-season performance of european union wheat forecasts during extreme impacts. *Sci. Rep.* 8, 15420. doi: 10.1038/s41598-018-33688-1

van Ginkel, M., Calhoun, D., Gebeyehu, G., Miranda, A., Tian-you, C., Lara, R. P., et al. (1998). Plant traits related to yield of wheat in early, late, or continuous drought conditions. *Euphytica* 100, 109–121. doi: 10.1023/A:1018364208370

Verde, N., Mallinis, G., Tsakiri-Strati, M., Georgiadis, C., and Patias, P. (2018). Assessment of radiometric resolution impact on remote sensing data classification accuracy. *Remote Sens.* 10, 1267. doi: 10.3390/rs10081267

Virnodkar, S. S., Pachghare, V. K., Patil, V. C., and Jha, S. K. (2020). Remote sensing and machine learning for crop water stress determination in various crops: a critical review. *Precis. Agric.* 21, 1121–1155. doi: 10.1007/s11119-020-09711-9

Volpato, L., Pinto, F., González-Pérez, L., Thompson, I. G., Borém, A., Reynolds, M., et al. (2021). High throughput field phenotyping for plant height using uav-based rgb imagery in wheat breeding lines: Feasibility and validation. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.591587

Vong, C. N., Conway, L. S., Feng, A., Zhou, J., Kitchen, N. R., and Sudduth, K. A. (2022). Corn emergence uniformity estimation and mapping using uav imagery and deep learning. *Comput. Electron. Agric.* 198, 107008. doi: 10.1016/j.compag.2022.107008

Vukasovic, S., Alahmad, S., Christopher, J., Snowdon, R. J., Stahl, A., and Hickey, L. T. (2022). Dissecting the genetics of early vigour to design drought-adapted wheat. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.754439

Walter, A., Liebisch, F., and Hund, A. (2015). Plant phenotyping: from bean weighing to image analysis. *Plant Methods* 11, 14. doi: 10.1186/s13007-015-0056-8

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12, 787. doi: 10.1111/pbi.12183

Wang, F., Yang, M., Ma, L., Zhang, T., Qin, W., Li, W., et al. (2022). Estimation of above-ground biomass of winter wheat based on consumer-grade multi-spectral uav. *Remote Sens.* 14, 1251. doi: 10.3390/rs14051251

Wang, L., Zhou, X., Zhu, X., Dong, Z., and Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* 4, 212–219. doi: 10.1016/j.cj.2016.01.008

Warrington, N. M., Tilling, K., Howe, L. D., Paternoster, L., Pennell, C. E., Wu, Y. Y., et al. (2014). Robustness of the linear mixed effects model to error distribution assumptions and the consequences for genome-wide association studies. *Stat. Appl. Genet. Mol. Biol.* 13, 567-587. doi: 10.1515/sagmb-2013-0066

White, J. W., Andrade-Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt, T. A., Conley, M. M., et al. (2012). Field-based phenomics for plant genetics research. *Field Crops Res.* 133, 101–112. doi: 10.1016/j.fcr.2012.04.003

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* 99, 673–686. doi: 10.1198/016214504000000980

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. (2013). Pitfalls of predicting complex traits from snps. *Nat. Rev. Genet.* 14, 507–515. doi: 10.1038/nrg3457

Wu, L., Chang, Y., Wang, L., Wang, S., and Wu, J. (2021). Genome-wide association analysis of drought resistance based on seed germination vigor and germination rate at the bud stage in common bean. *Agron. J.* 113, 2980–2990. doi: 10.1002/agj2.20683

Wu, C., Niu, Z., Tang, Q., and Huang, W. (2009). Predicting vegetation water content in wheat using normalized difference water indices derived from ground measurements. *J. Plant Res.* 122, 317–326. doi: 10.1007/s10265-009-0215-y

Xiong, P., Schnake, T., Montavon, G., Müller, K.-R., and Nakajima, S. (2022). Efficient computation of higher-order subgraph attribution via message passing. *Proc. Mach. Learn. Res.* 162, 24478–24495.

Xu, J., Lowe, C., Hernandez-Leon, S. G., Dreisigacker, S., Reynolds, M. P., Valenzuela-Soto, E. M., et al. (2022). The effects of brief heat during early booting on reproductive, developmental, and chlorophyll physiological performance in common wheat (triticum aestivum l.). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.886541

Xue, J., and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *J. Sensors* 2017, 1–17. doi: 10.1155/2017/1353691

Yadav, H. (2017). Effects of rain shelter or simulated rain during grain filling and maturation on subsequent wheat grain quality in the uk. *J. Agric. Sci.* 155, 300–316. doi: 10.1017/S0021859616000411

Yang, R. C. (2010). Towards understanding and use of mixed-model analysis of agricultural experiments. *Can. J. Plant Sci.* 90, 605–627. doi: 10.4141/CJPS10049

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008

Yang, Z., Rao, M., Elliott, N., Kindler, S., and Popham, T. (2009). Differentiating stress induced by greenbugs and Russian wheat aphids in wheat using remote sensing. *Comput. Electron. Agric.* 67, 64–70. doi: 10.1016/j.compag.2009.03.003

Yang, X., Song, Z., King, I., and Xu, Z. (2021). A survey on deep semi-supervised learning. *IEEE Trans. Knowledge Data Eng.* 35, 8934–8954. doi: 10.1109/TKDE.2022.3220219

Yazdani, A., Yazdani, A., Mendez-Giraldez, R., Samiei, A., Kosorok, M. R., and Schaid, D. J. (2022). From classical mendelian randomization to causal networks for systematic integration of multi-omics. *Front. Genet.* 13. doi: 10.3389/fgene.2022.990486

Yoosefzadeh-Najafabadi, M., Eskandari, M., Belzile, F., and Torkamaneh, D. (2022). Genome-wide association study statistical models: A review. *Methods Mol. Biol.* 2481, 43–62. doi: 10.1007/978-1-0716-2237-74/COVER

Yu, H., Zhang, Q., Sun, P., and Song, C. (2018). Impact of droughts on winter wheat yield in different growth stages during 2001–2016 in eastern China. *Int. J. Disaster Risk Sci.* 9, 376–391. doi: 10.1007/s13753-018-0187-4

Yuan, L., Bao, Z., Zhang, H., Zhang, Y., and Liang, X. (2017). Habitat monitoring to evaluate crop disease and pest distributions based on multi-source satellite remote sensing imagery. *Optik* 145, 66–73. doi: 10.1016/j.ijleo.2017.06.071

Yuan, W., Li, J., Bhatta, M., Shi, Y., Baenziger, P., and Ge, Y. (2018). Wheat height estimation using lidar in comparison to ultrasonic sensor and uas. *Sensors* 18, 3731. doi: 10.3390/s18113731

Zhang, Y.-M., Jia, Z., and Dunwell, J. M. (2019). [dataset] editorial: The applications of new multilocus gwas methodologies in the genetic dissection of complex traits. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00100

Zhang, J., Naik, H. S., Assefa, T., Sarkar, S., Reddy, R. V., Singh, A., et al. (2017). Computer vision and machine learning for robust phenotyping in genome-wide studies. *Sci. Rep.* 7, 1–11. doi: 10.1038/srep44048

Zhang, Q., Zhang, Q., and Jensen, J. (2022). Association studies and genomic prediction for genetic improvements in agriculture. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.904230

Zhao, Y., Potgieter, A. B., Zhang, M., Wu, B., and Hammer, G. L. (2020). Predicting wheat yield at the field scale by combining high-resolution sentinel-2 satellite imagery and crop modelling. *Remote Sens.* 12, 1024. doi: 10.3390/rs12061024

Zhao, J., Sun, L., Gao, H., Hu, M., Mu, L., Cheng, X., et al. (2023). Genome-wide association study of yield-related traits in common wheat (triticum aestivum l.) under normal and drought treatment conditions. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1098560

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zhou, X., Kono, Y., Win, A., Matsui, T., and Tanaka, T. S. T. (2021a). Predicting within-field variability in grain yield and protein content of winter wheat using uav-based multispectral imagery and machine learning approaches. *Plant Production Sci.* 24, 137–151. doi: 10.1080/1343943X.2020.1819165

Zhou, Z., Majeed, Y., and Naranjo, G. D. (2021b). [dataset] assessment for crop water stress with infrared thermal imagery in precision agriculture: A review and future prospects for deep learning applications. *Comput. Electron. Agric.* 182, 106019. doi: 10.1016/j.compag.2021.106019

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* 44, 821–824. doi: 10.1038/ng.2310

Zhou, L., Xiao, Q., Taha, M. F., Xu, C., and Zhang, C. (2023). Phenotypic analysis of diseased plant leaves using supervised and weakly supervised deep learning. *Plant Phenomics* 5, 0022. doi: 10.34133/plantphenomics.0022

Zhou, X., Zhang, J., Chen, D., Huang, Y., Kong, W., Yuan, L., et al. (2020). Assessment of leaf chlorophyll content models for winter wheat using landsat-8 multispectral remote sensing data. *Remote Sens.* 12, 2574. doi: 10.3390/rs12162574

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1. doi: 10.3835/plantgenome2008.02.0089

Zhu, W., Sun, Z., Huang, Y., Lai, J., Li, J., Zhang, J., et al. (2019). Improving field-scale wheat lai retrieval based on uav remote-sensing observations and optimized vi-luts. *Remote Sens.* 11, 2456. doi: 10.3390/rs11202456

Zhu, D., Wang, C., Pang, B., Shan, F., Wu, Q., and Zhao, C. (2012). Identification of wheat cultivars based on the hyperspectral image of single seed. *J. Nanoelectronics Optoelectronics* 7, 167–172. doi: 10.1166/jno.2012.1243

Zohary, D., Hopf, M., and Weiss, E. (2012). *Domestication of Plants in the Old World* (Oxford, UK: Oxford University Press). doi: 10.1093/acprof:osobl/9780199549061.001.0001

Glossary

AI	Artificial Intelligence
AM	Association Mapping
B	Blue band
CCCI	Canopy Chlorophyll Content Index
CMIP	Coupled Model Intercomparison Project
CNN	Convolutional Neural Network
CWSI	Crop Water Stress Index
DT	Decision Tree
EVI	Enhanced Vegetation Index
G	Green band
GIS	Geographic Information System
GLM	Generalised Linear Model
GNDVI	Green Normalised Difference Vegetation Index
GNN	Graph Neural Network
GS	Genomic Selection
GWAS	Genome-wide Association Studies
L	Canopy background adjustment factor
LAI	Leaf Area Index
LIME	Local Interpretable Model-agnostic Explanation
LM	Linkage Mapping
LMM	Linear Mixed Model
LRP	Layer-wise Relevance Propagation
LWIR	Long-wave Infrared
MAS	Marker-assisted Selection
ML	Machine Learning
MSAVI	Modified Soil Adjusted Vegetation Index
MTA	Marker-Trait Association
NDRE	Normalised Difference Red Edge
NDVI	Normalised Difference Vegetation Index
NDWI	Normalised Difference Water Index
NIR	Near-infrared
NN	Neural Network
R	Red band
SAVI	Soil-adjusted Vegetation Index
SHAP	Shapley Additive Explanation
SNP	Single Nucleotide Polymorphism
SVM	Support Vector Machine
SWIR	Shortwave Infrared
TCARI	Transformed Chlorophyll Absorption in Reflectance Index
UAV	Unmanned Aerial Vehicle
XAI	Explainable Artificial Intelligence.

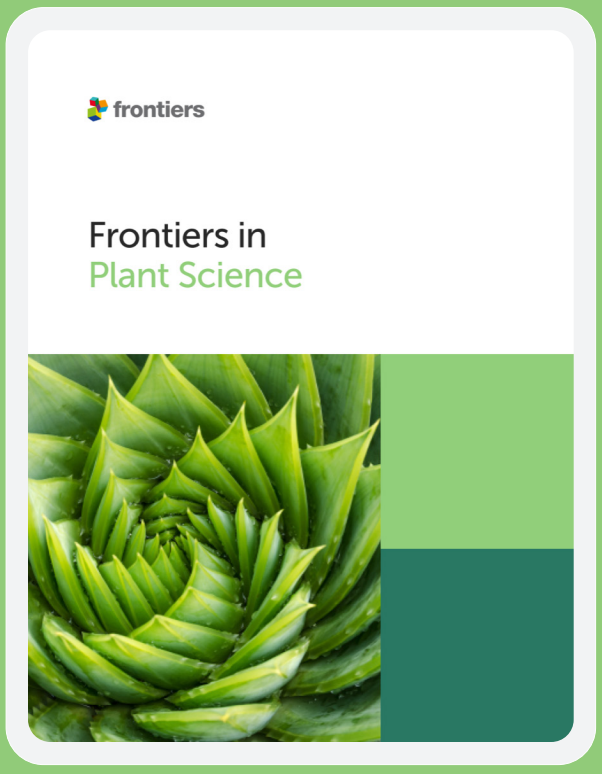
Frontiers in Plant Science

Cultivates the science of plant biology and its applications

The most cited plant science journal, which advances our understanding of plant biology for sustainable food security, functional ecosystems and human health.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact