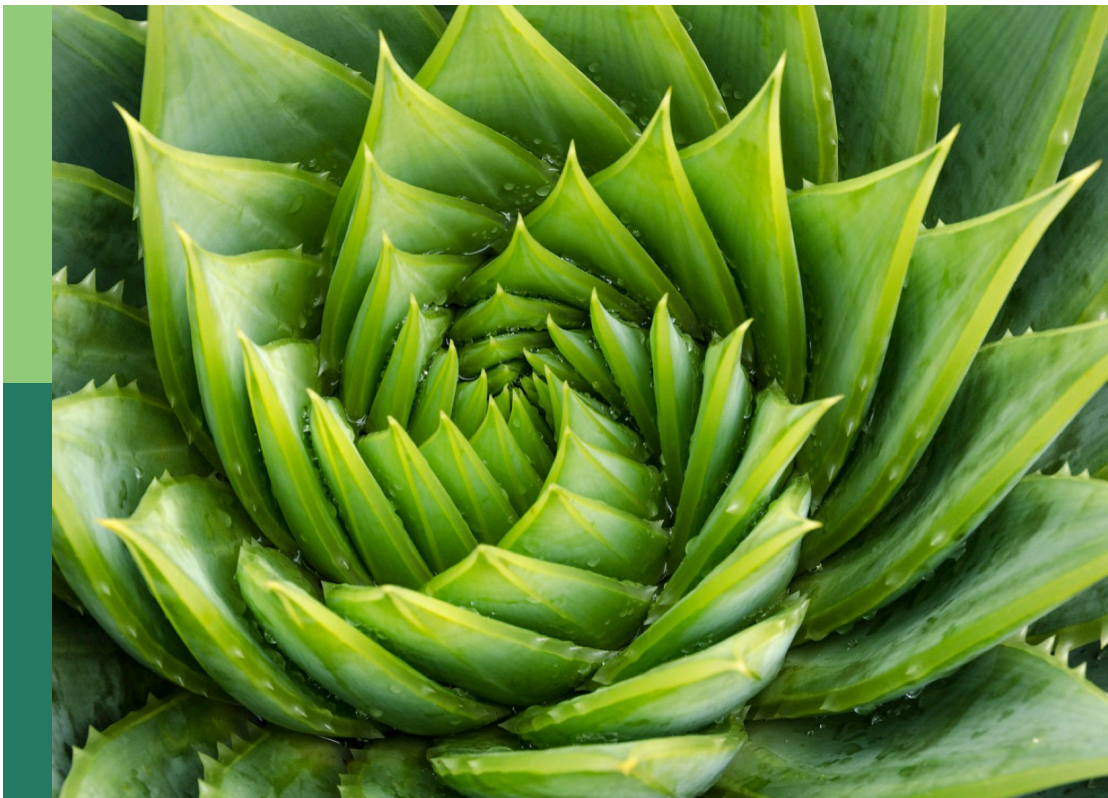
frontiers | Research Topics

Plant genotyping: From traditional markers to modern technologies

Edited by
Yuri Shavrukov, Patricio Hinrichsen and Satoshi Watanabe

Published in
Frontiers in Plant Science



ISSN 1664-8714
ISBN 978-2-8325-4891-2
DOI 10.3389/978-2-8325-4891-2

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

Plant genotyping: From traditional markers to modern technologies

Topic editors

Yuri Shavrukov — Flinders University, Australia

Patricio Hinrichsen — Agricultural Research Institute, Chile

Satoshi Watanabe — Saga University, Japan

Citation

Shavrukov, Y., Hinrichsen, P., Watanabe, S., eds. (2024). *Plant genotyping: From traditional markers to modern technologies*. Lausanne: Frontiers Media SA.

doi: 10.3389/978-2-8325-4891-2

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 30 April 2024
DOI 10.3389/fpls.2024.1419798

OPEN ACCESS

EDITED AND REVIEWED BY
Diego Rubiales,
Spanish National Research Council (CSIC), Spain

*CORRESPONDENCE
Yuri Shavrukov
yuri.shavrukov@flinders.edu.au
Patricio Hinrichsen
phinrichsen@inia.cl
Satoshi Watanabe
nabemame@cc.saga-u.ac.jp

RECEIVED 18 April 2024
ACCEPTED 23 April 2024
PUBLISHED 30 April 2024

CITATION
Shavrukov Y, Hinrichsen P and Watanabe S (2024) Editorial: Plant genotyping: from traditional markers to modern technologies. *Front. Plant Sci.* 15:1419798. doi: 10.3389/fpls.2024.1419798

Editorial: Plant genotyping: from traditional markers to modern technologies

Yuri Shavrukov[1]*, Patricio Hinrichsen[2]* and Satoshi Watanabe[3]*

[1]College of Science and Engineering (Biological Sciences), Flinders University, Adelaide, SA, Australia, [2]Instituto de Investigaciones Agropecuarias, INIA-La Platina, Santiago, Chile, [3]Faculty of Agriculture, Saga University, Saga, Japan

KEYWORDS

plant genotyping, traditional markers, modern technology, genome-wide association study (GWAS), simple sequence repeats (SSR), single nucleotide polymorphism (SNP)

Editorial on the Research Topic
Plant genotyping: from traditional markers to modern technologies

Unlike external plant traits, the naked human eye cannot distinguish the genotypes that comprise the underlying genetic material responsible for these phenotypic traits. To make genotypes accessible for research and further understanding and use in plant breeding and related topics, various genotyping methods have become available. Plant genotyping began with quite complex methods based on the direct hybridization of DNA fragments using labelled probes to identify specific genes, which required large quantities of target DNA (as in the case of Restriction fragment length polymorphism, or RFLP). After some years, they evolved into a large series of relatively simpler and cheaper PCR-based methods. These latter reached a peak with very polymorphic and straightforward markers, like microsatellites or SSR (Simple sequence repeats), which were then followed by DNA sequencing and fragment analysis, PCR and qPCR, allele-specific molecular probes and primers, and today's modern and advanced microchip-DNA technology involving hundreds to thousands of simultaneous reactions.

The current status of our knowledge and progress in plant genotyping was updated in this Research Topic, where we have detailed the available methods and technologies used to target various genes of interest in different plant species. A wide and diverse range of areas were covered and addressed in our Research Topic, from traditional molecular markers to modern microarray technologies. Various scientific approaches and research ideas were incorporated, all aimed at achieving a better understanding of and practical application of plant genotyping. This has led to the resulting 14 published papers that follow.

As mentioned above, SSR markers are a simple, versatile, and straightforward molecular tool for plant genotyping. Yin et al. used SSR markers for practical identification and distinctness testing of a non-heading Chinese cabbage (*Brassica campestris* ssp. *chinensis* Makino). This is a very important test that establishes distinctness, uniformity, and stability (DUS), which are essential factors required for the granting of plant variety rights (PVRs). The authors tested 287 SSR markers for genotyping of 423 non-heading Chinese cabbage varieties, and they used four fluorescent dyes, FAM, HEX, TAMRA and ROX, for the labelling of forward primers. Importantly, two methods

were used for scoring, polyacrylamide gel electrophoresis (PAGE) and fluorescence capillary electrophoresis. The resulting 23 core SSR markers were finally selected, enabling perfect genotyping of the majority of the studied non-heading Chinese cabbage varieties. Therefore, a combination of SSR genotyping with simple morphological markers in a field trial provided more accurate and efficient identification of the varieties for the distinctness test. Based on clustering analysis, the authors designated 423 varieties into three Clades and nice coloured photos illustrated their clear distinctness. This is one of the best examples of the simple and elegant application of plant genotyping using SSR markers in crops like non-heading Chinese cabbage.

An oat (*Avena sativa* L.) germplasm collection (132 cultivars and pure lines) with diverse origins was described by Mathias-Ramwell et al. for phenotype and genotype characterization within a Chilean breeding program. Specifically in Chile, a single cultivar (Supernova-INIA) is predominant, covering over 90% of the oat cultivated area. Therefore, this has forced the development of new oat varieties adapted to the changed climate, which is severely affecting the Southern part of Chile. This study combined the evaluation of 28 phenotypic traits and genotyping with 14 SSR markers that were previously reported as informative in oat. The studied oat germplasm collection exhibited a high phenotypic diversity (H' = 0.68) and grouped into three clusters. This result differed from the SSR-based Structure analysis indicating for the existence of two sub-populations with low genetic distance (0.24), despite moderate (He = 0.58) average genetic diversity. In summary, the combination of both phenotypic data and SSR-based genotyping supported the possibility to obtain genetic gain in the medium to short term in this breeding effort, opening the opportunity for improved oat germplasm materials.

Semalaiyappan et al. focused on pearl millet [*Pennisetum glaucum* (L.) R. Br.; syn. *Cenchrus americanus* (L.) Morrone], a strategic climate-resilient C4 crop and an important staple food in Asia and Africa. The authors retrieved 4K SNPs from 925 whole-genome sequences and carried out genotyping of 373 genetically diverse pearl millet inbred lines. Their genotyping of the SNP panel exhibited a uniform distribution across the entire genome. All studied accessions were effectively designated and differentiated using the SNP panel into two major groups (B and R lines) based on the genetic diversity analysis. The studied 4K SNP panel was reported as very useful for various genomics and molecular breeding applications in pearl millet, including mapping of agronomically important traits and genomic selection.

Genotyping of trees represents a very complicated process and is usually carried out on individuals established and grown over a very long time-frame. However, Wu et al. carried out genetic analysis of 69 parents and 1,793 third-generation offspring (ramets) in the seeds of orchard Chinese fir [*Cunninghamia lanceolata* (Lamb.) Hook]. This was very extensive research involving both morphological and molecular analyses. The authors used traditional SSR markers for plant genotyping to study the mating system and flowering phenology of trees. The SSR genotyping was based on fluorescent labels, FAM or HEX dyes, attached to forward primers. This approach is well known and widely used for plant genotyping, and it was very suitable for this study of Chinese fir trees. The results described genetic co-ancestry among parental genotypes that was detected in the third generation of ramets genotypes. Effective pollination (68.1%) occurred within 50 m, and it was successful if about 30% of male and female flowers overlapped in their flowering. It is important to emphasize that such an accurate and delicate study was achievable through plant genotyping using SSR markers.

Another study of a forest species was presented by Yan et al. The authors reported on the application of SSR markers for genotyping, analysis of genetic diversity and population structure in a collection of 161 Korean pine clones (*Pinus koraiensis* Siebold & Zucc.), originating from seven populations in Northeast China. A set of 77 alleles derived from 11 SSRs would at first seem very small but this was sufficient to accurately distinguish each clone. However, a rather low genetic diversity was exhibited among different populations, but diversity was higher within each studied population, explaining 98% of the total observed variation. This is a very unusual result for genotyping of Korean pine populations. Moreover, only one population, from Lushiuhe, was isolated and differentiated clearly. The set of 11 SSR markers used was proposed as a fingerprinting tool able to identify any specimen of Korean pine, and this final result can be potentially used for the breeding of this species.

Modern high-throughput genotyping microarrays provide thousands of simultaneous reactions, and Ding et al. presented a report on the successful application of 55K SNP microarrays for analysis of time to heading and maturity in a diverse group of 239 bread wheat accessions (*Triticum aestivum* L.). Starting from genome-wide association study (GWAS), the authors carried out three-year experiments in four environments. For genotyping, 16,649 high-quality SNP markers were selected and in the results of GWAS, 238 and 55 SNP markers were found to be strongly associated with time to heading and maturity, respectively. Finally, the authors identified only nine marker-trait associations in different environments with highest scores across the entire group of studied wheat genotypes. This resulted in nine SNPs in the most promising candidate genes controlling traits for time to heading and maturity in bread wheat. Many genes are involved in the control of such important traits as heading and maturity, and the authors discussed functions of these candidate genes in the paper: Zinc transporter and Zinc finger family protein, Glycosyltransferase and S-acyltransferase, F-box protein and Cytochrome P450, Calcium-dependent protein kinase and Photosystem II stability/assembly factor, and Cytokinin phosphoribohydrolase.

A genome-wide association study of sorghum was carried out by Wang et al. focusing on plant colour of sorghum [*Sorghum bicolor* (L.) Moench], which influences various traits such as seed colour as well as disease resistance and phytoalexin production. Using a sorghum mini-core collection, the authors assessed the colour of leaf sheaths and blades across three environments and conducted genome-wide association mapping with 6,094,317 SNP markers. Eight QTLs were identified and linked to plant colour, containing up-to 1-3 candidate genes each. These findings offer insights for the application of plant genotyping for plant colour development and in sorghum molecular breeding.

Two studies of maize (*Zea mays* L.) populations were based on the Iowa Stiff Stalk Synthetic (BSSS) germplasm stock. In the first

paper, Ledesma et al. described the molecular characterization of the collection of DH lines derived from the unselected BSSS population (C0) and those after 17 cycles of reciprocal recurrent selection in BSSS (C17). The progenies of a hybrid population between C0 and C17 were genotyped with a set of 24,885 SNP markers distributed among 10 maize chromosomes for evaluation of their genetic variability. The authors also studied the possible loss of genetic diversity during the recurrent selection process from C0 to C17. The reported results confirmed a net loss of variability with the degree of differentiation between C0 and C17 DH groups. The different contribution of the progenitors of DH lines derived from C0, C17 or their hybrid was mostly explained by genome-wide genetic drift. Additionally, complementary to allelic selection occurred during the reciprocal recurrent breeding supported by phenotype analysis data.

The continuation of the previous study with maize was reported by Ledesma et al. The authors applied GWAS for maize plant architecture traits, which were modified during the selection of BSSS populations with a very big impact on grain moisture and yield, root and stock lodging. Using the same approach, the authors compared phenotypes and genotypes of DH lines derived from BSSS recurrent selection. It included C17 DH, reciprocal recurrent selection (R) and from their hybrid. Plant phenotypes and studied agronomic traits as well as identified genes or genomic regions were associated with modifications in the plant architecture. Additionally, plant density and grain yield traits, including flowering time and time from anthesis to silking, showed high heritability and were more common for the BSSS(R)C17 DH lines. Finally, a considerable number of SNPs were identified in the genetic regions with promising candidate genes associated with plant architecture traits using the entire set of DH lines. Therefore, the genetic basis of the studied traits can be elucidated for marker-assisted selection schemes in maize breeding in future.

The effect of nitrogen fertilization levels on three related traits in maize (plant height, grain yield, and time from anthesis to silking) was explored by Sanchez et al., where phenotypic analysis was combined with GWAS for nitrogen use efficiency (NUE). For this purpose, 181 double haploid (DH) maize lines were studied using GWAS with 62,077 SNPs for plant genotyping. For three studied traits, data were collected from conditions of high or low nitrogen, under three environments, for both *per se* and testcross trials. Interestingly, significant genetic variation was observed among the DH lines and their respective testcrosses, using three GWAS models. Additionally, some testcrosses from exotic introgression lines were superior compared to the check hybrid. Finally, some SNPs were associated with agronomic traits under both high and low nitrogen. At the same time, these SNPs belonged to gene models and were related to stress response and nitrogen metabolism. In summary, this SNP-based GWAS analysis revealed the existence of several promising alleles in the maize germplasm panel with genes controlling key agronomic traits.

Genotyping by sequencing (GBS) is another approach for high-throughput plant genotyping. Lu et al. studied resistance of cabbage lines (*Brassica oleracea* L. var. *capitate*) to black rot disease (*Xanthomonas campestris* pv. *campestris*). The authors used GBS for QTL analysis of resistance in the $F_{2:3}$ hybrid population from a cross between resistant (BR155) and susceptible (SC31) parents. The genetic map was established with 7,940 SNP markers, and QTL analysis was carried out for disease resistance in 126 hybrid progenies over three seasons. In the results, the authors reported about seven identified QTLs with only one major QTL, qCaBR1, in chromosome C06. In the genetic interval of the major QTL 96 genes were annotated, but only eight of these genes showed responses to biotic and pathogenic factors. These candidate genes are listed as follows: Chorismate mutase, ß-1,4-N-acetylglucosaminyltransferase, Ethylene receptor, Plastid movement impaired, DNA ligase, Leucine-rich repeat protein kinase, RNA-binding family protein, and Early-responsive to dehydration protein.

Kiwifruit (*Actinidia chinensis* var. *chinensis*) can be attacked by one of the worst all plagues, *Pseudomonas syringae* pv. *actinidiae* (Psa). Therefore, the development of resistant germplasm is always a priority in the breeding of this species. Indeed, seedlings of certain genotypes can be highly susceptible to this disease, reaching up to 100% mortality. Flay et al. approached the search for QTLs associated to resistance to Psa, using a Bulk segregant analysis (BSA) approach. For this purpose, the authors analysed the effect of removing plants with Psa symptoms on the total allele frequency in the produced incomplete-factorial-cross population. The genotype-distinct diploid parents were used in this population consisting of 28 F_1 families. Only surviving plants from the different families were selected, their DNA was bulked, and QTLs were identified along with their detection accuracy. In addition, each family was assigned to a single bulk grouping according to the genetic contribution of a separate parent to each family. Finally, 11 QTLs were identified based on the deviation of allelic frequencies in the surviving populations in two independent analyses. This information was based on SNPs derived from a 30× bulk sequencing analysis. The authors have used their findings to initiate the development of novel molecular markers applied to the selection of kiwi lines with Psa resistance.

The development of universal markers for plant mitochondrial genomes is challenging because of their variability in size, gene order and sequence conservation. Grosser et al. presented a very interesting report exploring genetic polymorphism in mitochondrial introns to distinguish plant species. This is a very novel and non-traditional approach for differential plant genotyping. The researchers tested PCR primer sets across different angiosperm species and found that amplicon length was much more polymorphic among genera but significantly less within genera. The authors confirmed their results in different plant species. This study emphasized the utility of genetic polymorphism in intron length in the mitochondrial genome across various plant species. The presented results were estimated as providing important and valuable tools for potential applications in evolutionary studies and molecular-genetic research.

Very different was a paper presented by Tran et al., describing a method for single copy transgene identification through qPCR using the example of transgenic rice (*Oryza sativa* L.). The authors established a qPCR protocol for the reference gene *OsSBE4*, encoding starch branching enzyme, and the *nos* terminator used in the transgenic construct. The data reported a near 100% accuracy for the method in distinguishing homozygous single-insert transgenic plants. This assay could be successfully applied to other transgenic rice plants that have the *nos*

terminator in their construct. The standard conditions for qPCR can be used with relatively inexpensive dyes, such as SYBR Green. Therefore, the suggested qPCR method could be cost-effective and suitable for lower budget laboratories that are involved in rice transgenic research, or even modified to test transgenics of other species through the selection of primers for a known reference gene that perform comparably to the *nos* primers. The genotyping approach presented in the paper can be targeted not only toward transgene copy number, but also can be used to detect duplication of indigenous genes.

In summary, all 14 papers published in this Research Topic deal with very different aspects of plant biology, ecology, molecular genetics, and genomics, covering different crops and plant species, and quite diverse experimental designs. However, all these papers are united under the single topic of plant genotyping using different types of molecular markers. Some authors used the more traditional SSR markers while others were interested in SNP studies using GWAS and GBS technologies. The last group of researchers were compelled to investigate genetic polymorphism of intron length in mitochondrial genome or methods for the identification of transgene or endogenous gene copy numbers. In this regard, all presented results for plant genotyping are important not only in advancing scientific progress, but for their practical application in crop breeding, supporting biodiversity and biosecurity, and the analysis of plant-derived products for use in food, medicine, or other industries.

Author contributions

YS: Conceptualization, Writing – original draft. PH: Conceptualization, Writing – original draft. SW: Conceptualization, Writing – original draft.

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.

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

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 October 2022
DOI 10.3389/fpls.2022.1042290

OPEN ACCESS

EDITED BY
Yuri Shavrukov,
Flinders University, Australia

REVIEWED BY
Milan Lstibůrek,
Czech University of Life Sciences Prague, Czechia
Alexandre Magno Sebbenn,
Instituto Florestal, Brazil

*CORRESPONDENCE
Aiguo Duan
duanag@caf.ac.cn
Jianguo Zhang
zhangjg@caf.ac.cn

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

RECEIVED 12 September 2022
ACCEPTED 04 October 2022
PUBLISHED 27 October 2022

CITATION
Wu H, Zhao S, Wang X, Duan A and Zhang J (2022) Mating pattern and pollen dispersal in an advanced generation seed orchard of *Cunninghamia lanceolata* (Lamb.) Hook. *Front. Plant Sci.* 13:1042290. doi: 10.3389/fpls.2022.1042290

Mating pattern and pollen dispersal in an advanced generation seed orchard of *Cunninghamia lanceolata* (Lamb.) Hook

Hanbin Wu[1], Shirong Zhao[2], Xihan Wang[2], Aiguo Duan[1,3]* and Jianguo Zhang[1,3]*

[1]State Key Laboratory of Tree Genetics and Breeding & Key Laboratory of Tree Breeding and Cultivation, National Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing, China, [2]State-owned Forestry Farm of Weimin, Shaowu, China, [3]Collaborative Innovation Center of Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China

Seed orchards represent the link between forest breeding and conifer production forests, and their mating patterns determine the genetic quality of seed orchard crops to a large extent. We genotyped the parental clones and their open pollination offspring in the third-generation seed orchard of Chinese fir using microsatellite markers and observed the synchronization of florescence in the seed orchard to understand the genetic diversity and mating structure of the seed orchard population. Genetic coancestry among parental clones was detected in the third generation seed orchard of Chinese fir, and the genetic diversity of the open-pollinated offspring was slightly higher than that of the parental clones. The external pollen contamination rate ranged from 10.1% to 33.7%, 80% of the offspring were produced by 44% of the parental clones in the orchard, and no evidence of selfing was found. We found that 68.1% of the effective pollination occurred within 50 m, and 19.9% of the effective pollination occurred in the nearest neighbors. We also found that successful mating requires about 30% of florescence overlap between males and females, and there was a significant positive correlation between male reproductive energy and male parental contribution. Our results provide a valuable reference for the management and design of advanced generation seed orchards.

KEYWORDS

Chinese fir, genetic diversity, mating pattern, pollen contamination, seed orchard

1 Introduction

The tree selection breeding program aims to improve the economic value of future forests by planting excellent tree species with the highest genetic gain and diversity (El-Kassaby, 1995). A forest seed orchard is a special artificial forest composed of excellent clones or families selected manually, built according to the design requirements, and intensively managed (Hodge and White, 1993). A seed orchard is one of the most effective ways for conifer breeding; not only can it provide a large number of high-quality seeds, but it can also serve as a breeding base for improved varieties. Hence, a seed orchard is an important link in the breeding system from low to high levels (Sheng, 2014). They represent vectors that connect breeding and afforestation activities through packaging gain and the diversity of genetically improved seed crops (El-Kassaby, 1992; El-Kassaby, 2000a; El-Kassaby, 2000b). To fulfill this role, seed orchards are expected to function as closed, perfect panmictic populations (Eriksson et al., 1973), an ideal scenario that is rarely met due to the commonly observed variations in reproductive success and phenology among parental clones (Chen et al., 2019; Bian et al., 2020) as well as external gene flow (pollen contamination) from the ambient environment (El-Kassaby and Ritland, 1986; El-Kassaby, 1995). Therefore, maintaining genetic diversity in forest breeding programs is an arduous task, reflecting the trade-off between genetic diversity and genetic gain (El-Kassaby, 1995).

The mating pattern and genetic diversity level of seed orchards largely determine their adaptability, biological and abiotic resistance, and sustainability (Grattapaglia et al., 2014; Chen et al., 2018). The mating pattern of a seed orchard is mainly caused by the variation of female and male fertility between parents (Nielsen and Hansen, 2011), the difference in reproductive success rate (Torimaru et al., 2012; Funda et al., 2016), the synchronization of the flowering period (Li et al., 2011; Zhang et al., 2016; Muñoz-Gutiérrez et al., 2020), and the difference in pollen competitiveness (Nikkanen et al., 2000; Aronen et al., 2002). However, mating patterns have become more complex due to their expected co-ancestry accumulation in advanced generation breeding programs and seed orchards (Yang et al., 2021). An uneven parental gametic contribution (Chen et al., 2018), inbreeding, and pollen contamination (Wei et al., 2015; Yang et al., 2017; Song et al., 2018) may affect the genetic diversity of offspring and potential seed yield and quality.

Chinese fir (*Cunninghamia lanceolata* (Lamb.) Hook) has been deforested and utilized for about 8000 years and cultivated for more than 2000 years in China (Sheng, 2014). According to the ninth inventory of China's forest resources, the area of Chinese fir plantation reached 9.87 million ha, and the volume reached 755 million m^3. Genetic improvement of Chinese fir began in the 1950s. The first generation seed orchard, the second generation seed orchard, and the third generation seed orchard have been established successively, and the multi-generation genetic improvement procedure (Shi, 1994) and technical regulations for the seed orchard construction of Chinese fir (Xu et al., 2013) have been formulated. The fourth generation breeding process has started. However, with the rapid development of Chinese fir seed orchard generation, studies on parental mating patterns and flowering synchronization in the seed orchard are lacking. In addition, due to the increasing proportion of improved seeds used in afforestation activities, knowledge of the genetic variability of the seed orchard is crucial. Therefore, the evaluation of the mating mode and flowering synchronization of the seed orchard has important guiding significance for evaluating the function of seed orchards, implementing scientific management, and improving the seed yield and quality of seed orchards.

Here, we conducted a genetic analysis of parental and offspring genotypes in the third-generation seed orchard of Chinese fir in Fujian Province, China. We unraveled the population's mating system and flowering phenology using SSR genetic markers. Our research results describe the genetic diversity and pedigree structure and depict the mating system and its influencing factors in advanced generation seed orchards. In addition, our research provides insights for future breeding plans.

2 Materials and methods

2.1 Site data

The experimental site is located in the third-generation seed orchard of Chinese fir at Weimin state-owned Forest Farm, Nanping City, Fujian Province, China (long. 117.68471°E, lat. 27.049212°N). The mean annual temperature is 18.3 °C, and the annual precipitation is 1882 mm. The seed orchard was established in 2010, covering an area of 6.4 ha. The seed orchard has 69 parents, with 1793 ramets. It consists of seven blocks with 5 m × 5 m spacing. The 69 parents were divided into two groups, A and B (i.e., 1–35 and 36–69), which were staggered in seven blocks. Each block was distributed with an adjusted random block design. The spacing between plants of the same clone was greater than 20 m, avoiding the adjacent distribution of ramets of the same clone. A pollen isolation zone (*Pinus massoniana* Lamb., *Phyllostachys edulis* (Carriere) J. Houzeau) greater than 250 m was set around the seed orchard. In November 2019, seeds were collected in blocks 3 and 4. The cones of 45 parent clones were collected (24 parent clones had no cones), and the location of the mother trees was marked. Seedlings were raised in containers at Wugong Mountain Forest Farm, Jiangxi Province, China (long. 114.247439°E, lat. 27.297688°N). In September 2021, young leaves of seedlings were collected for DNA extraction, and 2–16 individuals were collected from each family. A total of 20 family samples were collected, and 288 offspring individuals' buds were collected (Table S1).

2.2 Genotyping

The total DNA of seed orchard parents and offspring seedlings was extracted with a CTAB plant genomic DNA rapid extraction kit (Adlai, Beijing, China). The DNA quality and concentration were detected by NanoDrop 2000. Six dinucleotides and six trinucleotide SSR markers (Li et al., 2015; Wen et al., 2015) were used for genotyping (Table 1). The forward primer of each primer pair was labeled with one of two fluorescent dyes (i.e., FAM or HEX) (Sangon Biotech, Shanghai, China). Polymerase chain reaction (PCR) analysis was carried out in 15 μL reaction volume: 50 ng genomic DNA template, 0.6 μL 10uM F-primer (including fluorescent primer), 0.6 μL 10uM R-primer, 12.8 μL 1.1 × Golden Star T6 Super PCR Mix (TsingKe, Beijing, China). The cycling parameters are also referred to in (Wen et al., 2015) The PCR cycling conditions consisted of an initial denaturation step of 95°C for 5 min; followed by 35 cycles of denaturation at 94°C for 30 s, annealing for 30 s (depending on the annealing temperature of the primer used, see Table 1), extension at 72°C for 30 s; and final extension at 72°C for 10 min. The amplification products were separated on the ABI3730 DNA analyzer (Applied Biosystems) using GeneScan-500 (LIZ) (Applied Biosystems) as an internal size standard. Allele binning and genotyping were performed automatically with Genemapper 4.0 software (Applied Biosystems) and later manually checked.

2.3 Data analyses

2.3.1 Genetic diversity analyses

GenALEx 6.5.1 (Peakall and Smouse, 2012) was used to calculate the genetic diversity of parents and offspring by analyzing the following parameters: observed (Na) and effective (Ae) number of alleles per locus, observed (Ho) and expected (He) heterozygosity, the Shannon diversity index (I), and fixation index (F) with F = (He – Ho)/He. The CERVUS 3.0 software (Kalinowski et al., 2007) was used to calculate parameters of genetic diversity: polymorphic information content (PIC), frequency of null alleles (Null), and Hardy–Weinberg equilibrium (HW).

2.3.2 Paternity analysis

Paternity analysis was conducted using the CERVUS 3.0 software (Kalinowski et al., 2007). Ten thousand simulations

TABLE 1 Primer information and characteristics of 12 microsatellite markers were evaluated using all samples in the present study.

Locus	Primer Sequences (5' - 3')	Accession NO. (GenBank)	Repeat Motif	Anneal temperature (°C)	Range (bp)	Na	Ae	Ho	He	HW[1]	Reference
wx1	ATTATCCGAGGCAGATACGCAC	AB749572	$(GGA)_{10}$	56	340-361	7	2.34	0.56	0.57	NS	Wen et al., 2015
	CTTCTCCGTATTTGATCCATCGC	AB749573									
wx2	GAGCCGTGAAGAACGAAGGTCTC	AB749574	$(GAA)_{12}$	56	261-285	8	4.14	0.77	0.76	NS	Wen et al., 2015
	ACGATCGGATTGTCTCAGAAACG	AB749575									
wx2-3	GATCCTCTGGTACTTGGTGCCC	AB749556	$(AT)_9$	56	184-196	6	1.62	0.33	0.38	NS	Wen et al., 2015
	TGCAAAGTCATGTCATCTCTGGC	AB749557									
wx2-6	TGAATGGACTGCCACAAATTCC	AB749550	$(AG)_{11}$	56	287-311	13	3.36	0.66	0.70	NS	Wen et al., 2015
	TTCTTTGCAGGAAAGCCAACAAG	AB749551									
wx2-4	GGCTCGAGTTTGCATCTCACAC	AB749558	$(TC)_9$	56	230-238	5	3.07	0.68	0.67	NS	Wen et al., 2015
	CACATCCAATCCATACAGGAGGG	AB749559									
wx4	AATGCGACTTGCAAATTTCTGG	AB749582	$(AGA)_{10}$	56	241-262	7	1.61	0.39	0.38	NS	Wen et al., 2015
	CGAATTCCTCAATCACTTGGCTG	AB749583									
wx2-11	TGATCTTGGCATGTCAGTCTGG	AB749576	$(AT)_9$	56	129-137	5	2.47	0.58	0.60	NS	Wen et al., 2015
	TGTCTGTCTGCCTGCAGTTATGC	AB749577									
wx8	TCCAGGAGTCTGTGAATCCGAAG	AB749600	$(CTG)_9$	56	203-233	8	2.05	0.50	0.51	NS	Wen et al., 2015
	CAGTACCAATTCAACCCAGCAGC	AB749601									
wx2-8	CTTAAGATAGCAGCGGGAATGG	AB749562	$(CT)_{11}$	56	240-260	11	3.24	0.49	0.69	***	Wen et al., 2015
	CTTGCTCGATTTCTTGCATCTGG	AB749563									
wx7	TTTGGGACCTTATGGAGGTGGAG	AB749602	$(GGA)_9$	56	122-146	7	2.21	0.52	0.55	NS	Wen et al., 2015
	AAACCACCAGGTTGAGAAGCAGC	AB749603									
wx6	GGAGCCCTTAGAGTTACGGAG	AB749578	$(ATA)_9$	56	211-223	5	2.22	0.48	0.55	NS	Wen et al., 2015
	TGGGCTCCATTCATTTGTACTGC	AB749579									
SM13	TCGTGAGTTTCTTGGTCATTTCG	KF873004	$(AG)_8$	61	385-399	7	2.62	0.35	0.62	***	Li et al., 2015
	CATAAGGGTTTTCCCCACGTATA										

[1]NS, not significant; ***, significant at the 0.1% level.
Na, observed number of alleles; Ae, effective number of alleles; Ho, observed heterozygosity; He, expected heterozygosity; HW, Hardy–Weinberg equilibrium.

were performed with 95% (strict) and 80% (loose) confidence levels. The critical LOD score was obtained by analyzing the simulated materials. The candidate father ratio and mistyped genotyping were set to 0.85 and 0.01, respectively, and the minimum number of loci was seven. The principle of CERVUS paternity analysis is that if the difference between the most probable and the second-most probable paternity exceeds a specific threshold (estimated through the simulation stage), the parent is set as candidate paternity (Kalinowski et al., 2007). We estimate the pollen contamination based on the individuals that do not match the father in the paternal analysis. The lower limit of the pollen contamination was based on the specific allele (29/288) in offspring, and the upper limit was that all unmatched individuals are produced by the peripheral pollen (97/288).

We estimated the male effective population size (Ne) following Funda et al. (2008). In order to compare the effective population size between orchards' crops and offspring population, we also estimated Ne with the linkage disequilibrium (LD) method in the program LDNE (Waples and Do, 2008).

2.3.3 Florescence

The fixed plant observation method of Bian et al. (2020) was used to observe the flower amount and flowering period. Four standard ramets with medium growth were selected for each clone in the seed orchard. The sunny side of each crown was divided into upper, middle, and lower layers. A branch with a medium flower amount was selected in each layer. The development process of male and female flowers was observed in spring, once a day until the end of the flowering period. The number of male and female flowers of the whole plant was observed at the last flowering stage. The evaluation criteria for each period of flowering of a single cone are referenced by Chen et al. (2019) and Bian et al. (2020).

The synchronization index of florescence proposed by Askew and Blush (1990) was used to evaluate the overlap of florescence. The basic idea was to evaluate the overlapping degree of the receptive and pollination stages among different genotypes. The calculation formula is given in Eq. 1:

$$S_{ij} = \sum_{k=1}^{n} \frac{\min\left(M_{ki}\ ,\ P_{kj}\right)}{\sum_{k=1}^{n} \max\left(M_{ki}\ ,\ P_{kj}\right)}\ , \qquad (1)$$

where S_{ij} is the synchronization index of florescence of clone i for the male parent and clone j for the female parent, M_{ki} is the loose pollen ratio of male strobilus of the ith grafted clone on the kth day; P_{kj} is the ratio at which the jth grafted clone is in the fertile period on the kth day; n is the number of days from the earliest flowering to the latest ending of male and female strobilus. When the florescence of male and female parents completely overlapped, $S_{ij} = 1$; when there is no overlap at all, $S_{ij} = 0$; In case of partial overlap, then $0 < S_{ij} < 1$. It is worth noting that Chinese fir is a monoecious species, and the same clone can be used as both male and female parents,$S_{ij} \neq S_{ji}$

2.3.4 Co-ancestry

Co-ancestry between parental clones was estimated using the triadic likelihood estimator (TrioML) with the software COANCESTRY (Wang, 2011). TrioML is expected to produce T of zero for unrelated cultivars, ~0.25 for half-sibs, and ~0.5 for full-sibs (Wang, 2011). Using a relatively small number of loci and possible genotyping errors reduced our estimated reliable T, so we applied TrioML to six superior cultivars of Chinese fir of known parentage (six hybrid combinations). From these cultivars, the lowest T estimated from first-degree relatives (full siblings or parent-offspring) was 0.4511; we thus used $T \geq 0.4511$ as a cutoff to identify first-degree relatives. Similarly, the maximum T-value of seven parental clones from different sources selected in the three first-generation seed orchards was 0.1641. We believe that the combination with $T < 0.1641$ was not related. Although these values may represent the true pedigree relationship, it is also possible that other complex mating schemes, such as backcross between generations, produce T-values equivalent to first-degree relatives. The relatedness between parental clones was represented by a weightless and directionless network, plotted with the R package 'igraph' (Csardi and Nepusz, 2006).

3 Results

3.1 Genetic diversity of parental and offspring population

In the analysis of 12 pairs of primers, the observed number of alleles (Na) varied from 4 (wx2-4) to 9 (wx2-6) in third-generation seed orchard parental clones, with a mean value of 6.083 (Table S2). The mean effective number of alleles (Ae) was 2.615. Among loci, observed heterozygosity (Ho) and expected heterozygosity (He) ranged from 0.375 to 0.853 and from 0.397 to 0.793, with means of 0.528 and 0.572, respectively. Primer wx2-8 showed significant deviation from Hardy–Weinberg equilibrium. A total of 80 alleles were amplified from 12 microsatellite loci in 288 Chinese fir offspring, with an average of 6.667 per primer (Table S3). Ho and He values ranged from 0.439 to 0.835 and 0.412 to 0.741, with means of 0.614 and 0.574, respectively. Loci wx1, wx2, wx2-8, wx2-4 and SM13 significantly deviated from the Hardy–Weinberg equilibrium. Compared to the parental population, the observed number of alleles (Na), and the proportion of observed heterozygosity in the offspring displayed an increase(Ho=0.528 and He=0.572, Ho=0.614 and He=0.574).r

3.2 Co-ancestry between parental clones

We found that most parental clone pairs were uncorrelated (86.19%), and the estimated T was less than 0.1641 (Figure 1A). In fact, most of the combinations estimated T = 0 (56.01%) or

less than 0.1 (77.15%). In addition, 11.85% of the combinations were between 0.1641 and 0.4511 (Figure 1A), indicating that they are related but may not be first-degree relatives and may be half-sib. Finally, 1.96% of the combinations were identified as first-degree relatives, of which the estimated T of the three combinations was equal to one (Figure 2B), so we were led to believe that they are clones of each other. Overall, 46 (65.7%) of the third-generation seed orchard parental clones had at least one first-degree relative (Figure 2A); P18 and P19 had five close relatives, followed by P3 and P7, which had four first-degree relatives. Any parental clone of the seed orchard had at least three related parental clones in the seed orchard (Figure 3A); P3, P4, and P7 had 18 parental clones related to them (Figure 3B).

3.3 Paternal analysis of the open-pollinated offspring

The results of parental analysis of 288 offspring showed that no selfing was found, and the outcrossing rate was 100% (Table 2). There were 191 (66.3%) offspring that were matched to the paternal clones, the remaining 97 offspring individuals were not matched, and 29 individuals in the offspring population found alleles (Na) unique to the offspring that did not appear in the parental population, which indicated that the pollen contamination rate of the seed orchard was 10.1%–33.7%, the average pollen contamination of the family was 32.3% ± 3.6%. There were 58 (86.6%) parental clones that participated in pollination. The effective pollination times of clones ranged from 1 to 11 (Figure 4A). The parental clone called P49 had the most pollination times, with a contribution rate of 5.76%. We found that 44% of the paternal clones produced 80% of the offspring, which indicated that the contribution rate of paternal clones in the seed orchard was unbalanced (Figure 4B).

Male effective population of seed orchard offspring (♂Ne) was 37.57, and ♂Ne/N ratio was 0.54 (Table S4). If it is assumed that each pollen contamination involved a single father, the ♂Ne increased considerably (by 103%). The Ne of parents and offspring population in seed orchard estimated by LD method was 49.00 and 62.30 respectively.

3.4 Effective pollen transmission distance

The effective propagation of pollen among clones in the seed orchard was random, and there was no obvious specific direction for pollination. The distance between each mother tree and its corresponding pollen donor was 5–176 m, and the average propagation distance was 49 m (Table 3); 68.1% of the effective pollination was within 50 m, and 19.9% of the

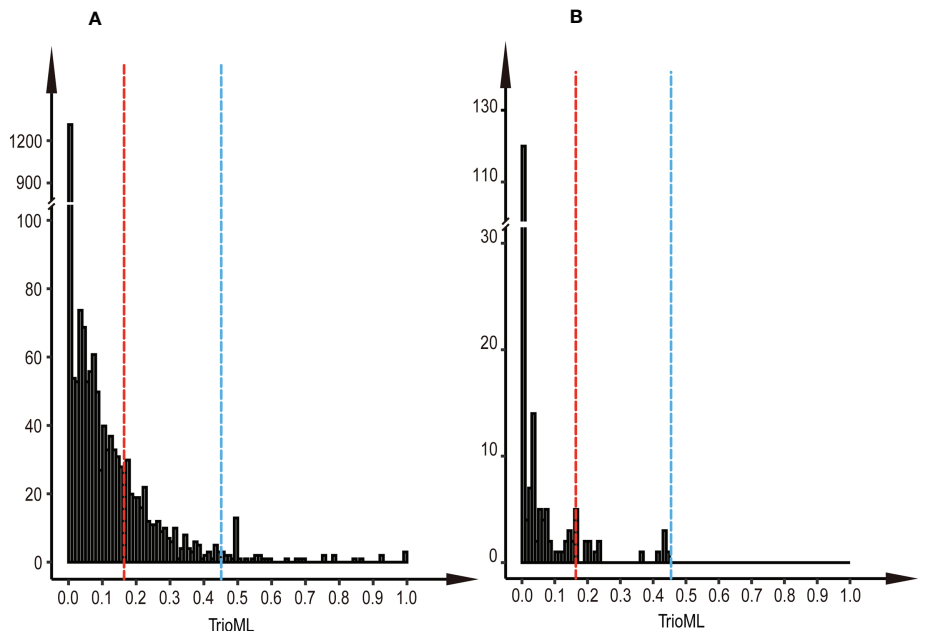


FIGURE 1
Histogram of paired TrioML values. **(A)** Histogram of paired TrioML values of parental clones in the third-generation seed orchard. **(B)** Histogram of paired TrioML values between parents with 191 cases of mating success. The red dotted line represents the unrelated cutoff of the parental clones (T = 0.1641), and the blue dotted line represents the threshold above which pairwise comparisons were considered first-degree relatives (T = 0.4511).

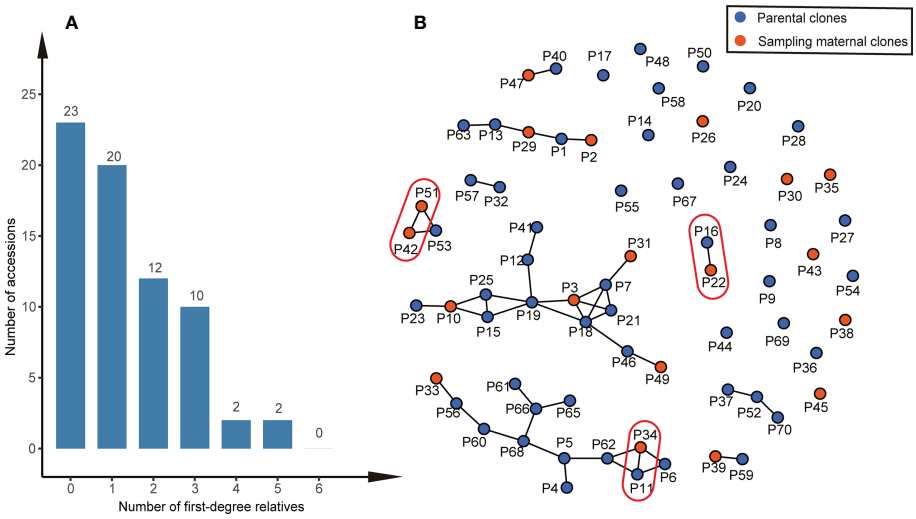


FIGURE 2
First-degree relationships within the Chinese fir germplasm collection. **(A)** Histogram of the number of first-degree relatives in the thirdgeneration seed orchard of Chinese fir. **(B)** A network of parental clones shows connections between first-degree relatives. The T (TrioML) of the parental clones in the red oval was one, and they were considered clones of one another.

effective pollination occurred in adjacent areas. All maternal clones showed substantial variation in mean pollination distance, with a mean coefficient of variation of pollination distance of 84.6% (Table 3).

3.5 Florescence synchronization

The synchronization index of florescence (*S*) for the Chinese fir parental clones was 0.48, with a minimum value of 0.045 and a maximum value of 1.00. Nine clones did not produce female flowers; consequently, their *S*-values were set to zero (red column in Figure 5). In the 191 cases of successful reproduction, the lowest *S*-value was 0.278 (dotted line in Figure 5). Excluding the clonal parents that did not produce female flowers, only 7% of the mating combinations in the seed orchard had an *S*-value lower than 0.278.

3.6 Mating success

The paternal contribution of clones was significantly different (Chi-square test, $p<0.001$), and the number of ramets

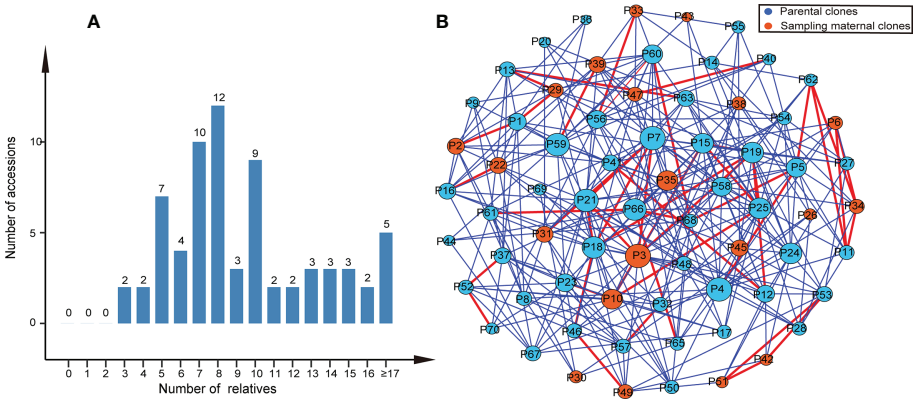


FIGURE 3
Relatives within the Chinese fir germplasm collection. **(A)** Histogram of the number of relatives in the third-generation seed orchard of Chinese fir. **(B)** A network of parental clones showing connections between relatives. The two circles connected by the red lines represent parental clone pairs were first-degree relatives, and the blue lines represent that parental clone pairs were related but probably not first-degree relatives. The larger the circle, the more parental clones were related to it.

10.3389/fpls.2022.1042290

TABLE 2 Paternity analysis for the 288 offspring.

Maternal clone	Number of offspring	Number of assigned offspring	Rate (%)	Number of assigned paternal clones	Assigned paternal diversity (%)	Outcrossing rate (%)
P10	15	10	66.67	5	50.00	100
P2	15	8	53.33	8	100.00	100
P22	15	14	93.33	12	85.71	100
P26	16	12	75.00	9	75.00	100
P29	15	7	46.67	6	85.71	100
P3	15	6	40.00	6	100.00	100
P30	15	13	86.67	9	69.23	100
P31	15	9	60.00	8	88.89	100
P33	15	9	60.00	5	55.56	100
P34	15	8	53.33	7	87.50	100
P35	15	11	73.33	8	72.73	100
P38	15	10	66.67	9	90.00	100
P39	15	11	73.33	7	63.64	100
P42	15	14	93.33	12	85.71	100
P43	15	11	73.33	11	100.00	100
P45	15	9	60.00	7	77.78	100
P47	15	9	66.67	7	77.78	100
P49	15	8	53.33	6	75.00	100
P51	15	10	66.67	8	80.00	100
P6	2	2	100.00	2	100.00	100
Total	288	191	—	152	—	—

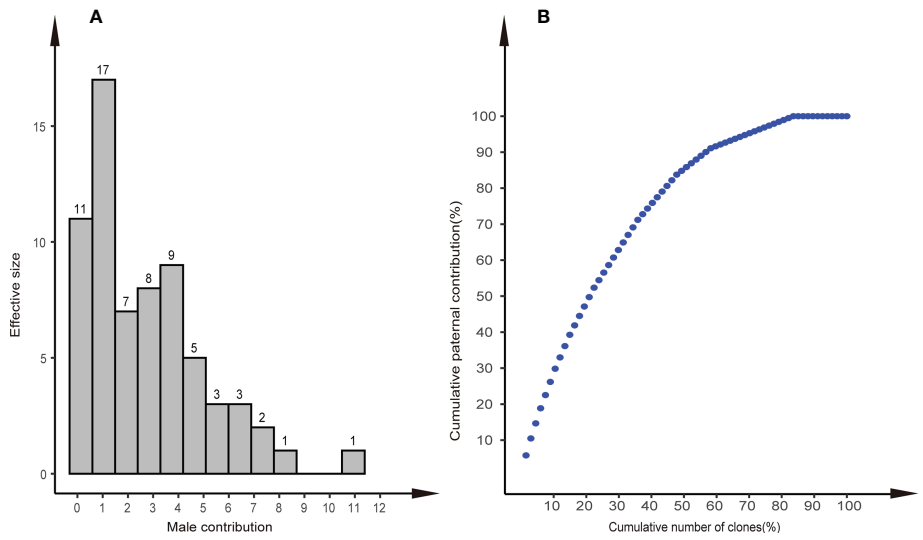


FIGURE 4
Distribution of paternal contributions and cumulative paternal contribution. **(A)** Distribution of paternal contributions of the 58 identified paternal clones in the Chinese fir clonal seed orchard. **(B)** Relationship between the cumulative number of clones (%) and cumulative paternal contribution (%) in the seed orchard.

TABLE 3 The difference in mean pollen transmission distance among female parent clones.

Maternal clone	Number of assigned paternal clones	Mean (m)	St. Error	CV (%)
P10	10	36.69	5.11	13.94
P2	8	42.62	49.23	115.51
P22	14	58.58	46.90	86.21
P26	12	32.44	30.71	104.98
P29	7	42.02	38.58	89.65
P3	6	72.99	72.78	99.70
P30	13	35.64	41.11	120.30
P31	9	36.44	36.43	100.21
P33	9	63.52	34.52	54.34
P34	8	79.14	42.51	60.86
P35	11	50.67	48.12	89.04
P38	10	55.89	44.63	90.22
P39	11	33.99	27.87	82.00
P42	14	40.33	41.18	102.11
P43	11	61.47	18.99	38.85
P45	9	22.80	36.22	158.86
P47	9	40.79	35.65	49.93
P49	8	53.14	46.18	85.67
P51	10	74.38	50.40	74.82
P6	2	95.49	28.97	30.34
Total	191	46.91	40.79	84.60

of clones explained 6.86% of the paternal contribution (Figure 6A). There was a significant difference in the number of flowers per clone among clones in the seed orchard (ANOVA, $p<0.001$; Table S5), and a positive correlation between the number of male flowers per clone and their paternal contribution rate ($r = 0.32$, $p = 0.0141$); their pollen yield explained 8.91% of the paternal contribution (Figure 6B). The *S*-value of mating success was not related to its paternal contribution (Figure 6C). There was a weak negative correlation between paternal contribution and T value of parents with mating success ($r = - 0.12$, $p = 0.082$) (Figure 6D). In addition, we found that the T-value of most mating parents (91.7%) was less than 0.1641 in 191 cases of mating success, which indicated that most parental clones of mating success were not related, and 8.3% of parents were related but did not reach the level of first-degree relatives (Figure 1B).

4 Discussion

4.1 Genetic diversity and co-ancestry

The main function of forest seed orchards is to produce a large number of genetically improved seeds without reducing genetic diversity (Chaloupková et al., 2019). For a long rotation species, the area where the species is planted should monitor the deployment of improved Germplasm to ensure that the level of genetic diversity is maintained and equivalent to the natural regeneration stand (Ruņģis et al., 2019). In this study, we compared the genetic diversity indexes of parental clones and offspring populations in the third-generation seed orchard of Chinese fir trees. The results show that the genetic diversity of parental clones and offspring was high, the mean observed heterozygosity (Ho) and expected heterozygosity (He) of parental clones were 0.528 and 0.572, and other coniferous seed orchard plants also had similar levels (*Larix kaempferi*, He = 0.525; *Larix olgensis*, He = 0.5833) (Wei et al., 2015; Chen et al., 2018). The expected heterozygosity (He) was slightly lower than that of the Chinese fir provenance population in the same province (He = 0.625) (Duan et al., 2017), which may be related to the population size and the generations of Chinese fir seed orchard. One study showed that the genetic diversity of the advanced generation breeding population gradually decreased as breeding progressed (Li et al., 2020). The Ho and He of the free pollinated offspring in the Chinese fir third-generation seed orchard were 0.614 and 0.575, respectively. The genetic diversity of the free-pollinated offspring was slightly higher than that of the parental clones. Similar reports have also been made on *Picea abies* (L.) Karst. (Sønstebø et al., 2018), *L. kaempferi* (Chen et al., 2018), and other tree species (Chaix et al., 2003; Yang et al., 2016). The genetic diversity of the progenies was higher than that of the parental clones, which was related to the introduction of new alleles and haplotypes by

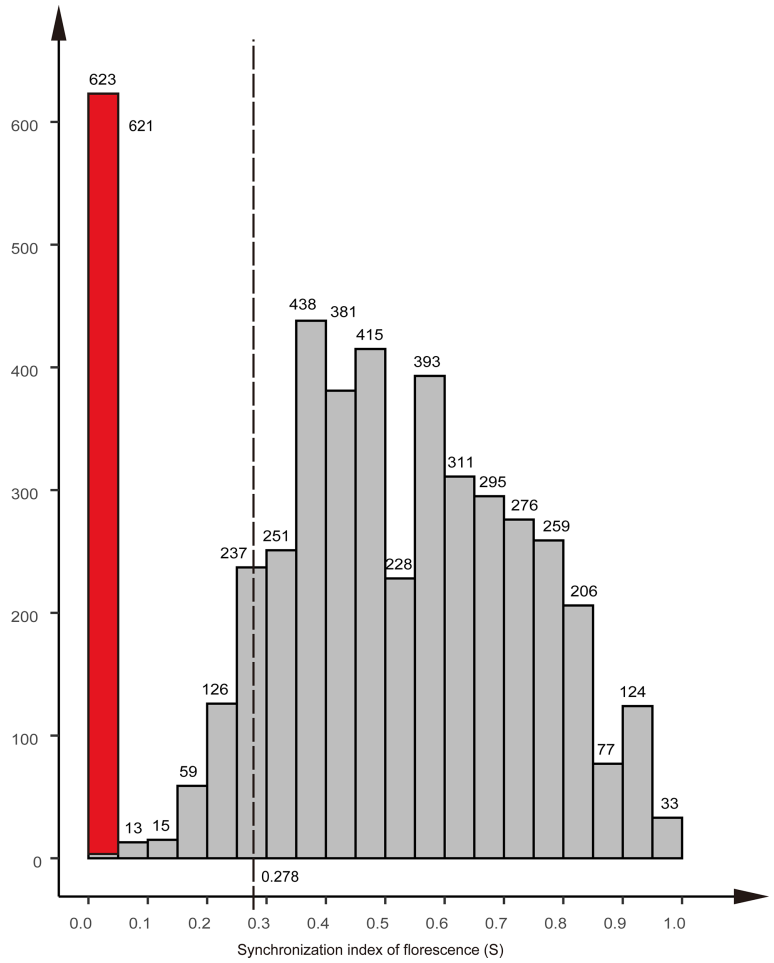


FIGURE 5
Distribution of the synchronization index of florescence (S) of the clones in the Chinese r clonal seed orchard. The red column indicates the nine clones that did not produce female flowers. equal to zero; The dotted line indicates the lowest S-value of mating success.

external pollen (Kess and El-Kassaby, 2015). In this study, the seed orchard had high pollen pollution (10.1%–33.7%).

Mating between relatives in the seed orchard population usually leads to the reduction of seed yield and the performance of inbreeding seedlings, which will continue throughout the tree life cycle (Woods and Heaman, 1989; Woods et al., 2001; Wang et al., 2004; Stoehr et al., 2014). In this study, genetic coancestry among parental clones was detected in the third generation seed orchard of Chinese fir, more than half (65.71%) of the parental clones had at least one first-degree relative, and any parental clone of the seed orchard had at least three related parental clones in the seed orchard. The build-up of co-ancestry had also been found in the advanced generation seed orchards of *Pinus tabuliformis* Carrière and *Pinus sylvestris* var. mongolica (Yang et al., 2020; Yang et al., 2021). This phenomenon of the build-up of co-ancestry was related to the determination of the progeny in the seed orchards and the selection of breeding populations. The families with the best performance often contributed more selected individuals than those with poor performance (Yang et al., 2021). The understanding of the parental relationship in seed orchards not only has important guiding significance for the breeding plan and process but also can be used as the basis for the spatial deployment design of parental clones in seed orchards.

4.2 Selfing and pollen contamination

Most trees are outbreeding species that have developed different mechanisms against selfing: dioecism, gametophytic self-incompatibility, the avoidance of contemporary flowering of male and female parts in hermaphrodite flowers, the separation of female and male flowers in different parts of the crown, or

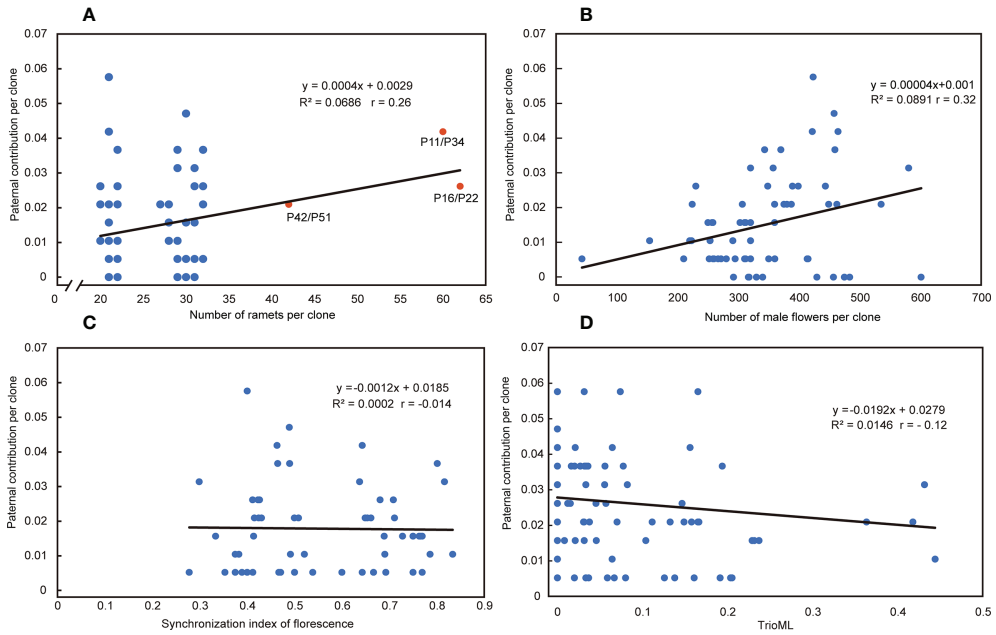


FIGURE 6
Relationship between reproductive factors and paternal contribution. **(A)** Relationship between the number of ramets per clone and paternal contribution per clone in the seed orchard. Red dots represent that pairwise comparisons were considered clones. **(B)** Relationship between the number of male flowers per clone and paternal contribution per clone in the seed orchard. **(C)** Relationship between the synchronization index of florescence (S) and the paternal contribution per clone in the seed orchard. **(D)** Relationship between the T (TrioML) and the paternal contribution per clone in the seed orchard. R2= R-Squared; r= Pearson correlation coefficient.

frequent abortion of embryos after selfing resulting in empty seeds (Liesebach et al., 2021). Selfing will cause recession, resulting in higher seedling mortality, etiolated seedlings, and poor individual vitality (Moriguchi et al., 2005). Therefore, evaluating and reducing the selfing rate of seed orchards in such ideal closed environments is very important.

It has been reported that the selfing rate of most conifer seed orchards is at low levels, with 5.1% for *Pinus sylvestris* L. (Funda et al., 2015), 0.45% for *Platycladus orientalis* (Huang et al., 2018), 1.85% for *L. olgensis* (Wei et al., 2015), 1.74%–6% for *P. abies* (L.) Karst. (Sønstebø et al., 2018); but there are exceptions, the selfing rate of a *Pseudotsuga menziesii* (Mirb.) Franco seed orchard is 12%–17% (Korecký and El-Kassaby, 2016; Song et al., 2018). Selfing was not detected in this study. On the one hand, it was related to the configuration of the seed orchard, and the distance between ramets of the same clone was more than 20 m; on the other hand, it is related to the tree species characteristics of Chinese fir (wind pollination), as female flowers in the phenological period mostly grow at the top of the trees (Zhang, 2005), and the seed orchard has auxiliary pollination measures. Similarly, selfing was not detected in the 1.5th generation seed orchard of red-heart Chinese fir (Chen et al., 2021). Previous studies have found that the self-pollination of Chinese fir can produce normally developed offspring (Xu et al., 2015; He et al., 2016), but all the cases of self-pollination were controlled pollination (only the pollen of its own clones). We speculated that the mating of open pollination of Chinese fir might be related to the affinity of pollen, as female flowers may preferentially choose the pollen of unrelated clones in the fertilization process. When the pollen of other clones is low or absent, plant-selective selfing provides reproductive security for the population. However, this conjecture about Chinese fir needs to be verified by further research. We also found low levels of inbreeding (Figure 6B). Inbreeding has generally been regarded as something to avoid in seed orchards, lesser relatedness between parents would lead to more successful fruiting, with inbreeding depression expressed only later in the established plantation (Mullin et al., 2019). It is generally recommended to manage the breeding population with separate sublines so that individual, unrelated seed orchard parents can be selected from each subline (McKeand and Bridgwater, 1998). This way focus on the gene pool of the seed orchard crop from the orchard. In an addition, Lindgren et al. (2009) recommended that under the constraint of overall genetic diversity (or number of states), the relatedness of pairs of seed-orchard parents were penalized to reduce the expected inbreeding.

Pollen contamination from outside the seed orchard tends to increase the genetic diversity of the seed orchard with a small number of parental clones (El-Kassaby and Ritland, 1986; Lindgren and Prescher, 2005), but it may also lead to the introduction of poor genetic material, so it should be prevented (Pakull et al., 2021). The mean pollen

contamination in the two seed orchards of *P. abies* (L.) Karst. was 22.9%, *P. menziesii* (Mirb.) Franco seed orchards reported pollen contamination rates of 10%–18%, and up to 28% under natural conditions (Sk Lai et al., 2010; Kess and El-Kassaby, 2015; Korecký and El-Kassaby, 2016; Song et al., 2018). In this study, even if there was an isolation zone far greater than the mean pollen transmission distance, the pollen contamination rate of the third-generation seed orchard of Chinese fir was between 10.1% and 33.7%; moreover, the unique allele (Na) of the offspring was also observed. We speculate that there was a large area of farmland between blocks 3-4 and 6-7 where the farmers planted Chinese fir on farmland after the farmland was abandoned, and these trees reached the stage of reproductive development. We believe that these Chinese firs were the source of pollen contamination donors.

Knowledge about the effective population size and pollen pollution is crucial to the establishment and management of seed orchards (Sønstebø et al., 2018). The effective population size is lower than the number of parents in the seed orchard (i.e., the number of parents), and the ratio of male effective population size to the census population size (♂Ne/N) was 0.54 in the seed orchard, which is consistent with most previous studies on the seed orchard (Funda and El-Kassaby, 2012; Sønstebø et al., 2018). The Ne estimated by LD method was larger than that estimated by paternity analysis (♂Ne), which is related to the high level of pollen contamination in the seed orchard, the pollen contamination increased the effective population size (Nikkanen and Ruotsalainen, 2000).

4.3 Paternal contribution and pollen transmission distance

It is generally believed that equal paternal contribution is very important for improving the genetic quality of seeds (El-Kassaby and Reynolds, 1990; Wheeler and Jech, 1992; Stoehr et al., 1998). Equal parental representativeness tends to produce the same gamete contribution as the parents in the seed orchard, thus reflecting their allele frequencies and meeting the basic quantitative genetic hypothesis of trait generation transmission (El-Kassaby and Sziklai, 1982). However, this is an ideal state based on the assumption of an equal reproductive yield of parents (male and female gametes) (Moriguchi et al., 2004). In this study, the paternal contribution rate of the third-generation Chinese fir seed orchard was in an unbalanced state, 80% of the gamete contribution comes from 44% of the parental population, and a similar proportion has been reported in other conifer seed orchards (Sk Lai et al., 2010; Song et al., 2018; Pakull et al., 2021; Heuchel et al., 2022). The reproductive success of the male parent was related to the number of ramets of the clone and the distribution of the male parent of the clone, Torimaru et al.(2012) found that the number of ramets of each clone was positively correlated with their paternal contribution in *P. sylvestris* seed orchard. In this study, we also found this positive correlation (r = 0.26), and the paternal contribution of clones was significantly different. The number of ramets of clones explained relatively little (6.86%) for the contribution of male parents. The mating success of male parents also depends on the pollen yield, flowering phenology, germination vigor of pollen grains, germination time, pollen tube growth rate, selective fertilization, and other traits (Aronen et al., 2002). Still, the major premise was that male and female florescence overlap. Our study showed that the mean *S*-value was 0.555 and the minimum *S*-value was 0.278 in 191 successful mating cases, which indicates that pollination can be completed so long as there is about a 30% overlap in florescence, even though there was no obvious correlation between the *S*-values and the paternal contribution rate (Figure 6C). It follows that the flowering phenology was not the main reason for the unbalanced paternal contribution. There was a significant difference in the number of male flowers among clones in the seed orchard ($p<0.001$) and a positive correlation between the number of male flowers per clone and their paternal contribution rate (r = 0.32). The pollen yield explained 8.91% of the contribution of male parents, noting that a higher proportion (~75%) of the more accurate index (pollen weight per clone) was used in the *P. sylvestris* L. seed orchard (Torimaru et al., 2012).

Knowledge of the effective pollination distance and mating system is important for the protection and management decisions of seed orchards and tree populations (Sork et al., 2002; Funda et al., 2008). In this study, the mean effective pollination distance of the third-generation seed orchard of Chinese fir was 47 m, and the longest distance was 176 m; 65.4% of the effective pollination occurred within 50 m, and 19.9% of the effective pollination occurred in the neighborhood, thus supporting the pollen transmission mode of short distance transmission. The pollination distance of Chinese fir maternal clones depends on their position in the seed orchard and the affinity between mating gametes. The wind direction (Feng et al., 2010), meteorological factors, (Huang et al., 2018) and management measures (such as auxiliary pollination) of the seed orchard during the flowering period will have an impact on the mattings. Pollen movement and embryo competition are complex processes. Previous studies on the pollen transmission of Chinese fir have shown that the pollen diffusion distance of Chinese fir is between 150 m and 600 m (Jiang et al., 1986; Chen, 1991; Chen et al., 1996). This gap may be due to different research methods. Previous studies mostly focused on direct observations of pollen density at different locations. The effectiveness and accuracy of this method were unstable and, in most cases, will not truly reflect the actual pollination and gene flow. The pollen dispersion law also depends on the population size, spatial distribution, and habitat characteristics (Ou, 2012).

4.4 Implications for the management and spatial design of a Chinese fir advanced generation seed orchard

Evaluation of mating patterns and pedigree structure of seed orchards can provide valuable information for scientific management and layout design of seed orchards. The limited number of marker loci used in this study, together with the difficulty of correctly scoring null alleles in these marker systems may have constrained assignment precision (Funda et al., 2015). SNP (single nucleotide polymorphism) analyses based on hundreds or thousands of markers have led to a stronger power of precise parentage reconstruction than microsatellite SSRs (Laucou et al., 2018). They have become more and more popular because they can be high throughput genotyping through next-generation sequencing (NGS) at a moderate cost (Zheng et al., 2019). The application of this technology in higher-generation seed orchard populations should be strengthened in future studies.

As the material for paternity analysis is the normal growth and development of the seed orchard's progeny population, most empty and astringent seeds (inactive) or seeds containing lethal alleles were produced by selfing, and the death of seeds at the germination stage was excluded (González-Martínez et al., 2001). Therefore, the actual selfing rate may be higher than reported in this study. The size limit of the offspring population may also ignore some successful mating cases, but the entire study simulates the production process of advanced generation seed orchards. From flowering and mating to the offspring plants with normal development, it can provide effective guidance for production practice.

Based on the universality of tree species distribution, seed orchards have been established in both marginal and central production areas. The mating systems of seed orchards are different under the trend of climate in different distribution areas and global climates. Therefore, a more comprehensive study is needed in different regions and years in future studies. Due to the annual and seasonal differences in the reproductive capacity of parents in seed orchards (Funda et al., 2009; Muñoz-Gutiérrez et al., 2020), continuous observation of mating and florescence in seed orchards is greatly significant to the management and construction of seed orchards and the formulation of breeding policies. However, continuous observation is a laborious and costly study. The individual contributions of parents in *P. abies* (L.) Karst. seed orchard was well correlated between the two studied years (Sønstebø et al., 2018). In this study, the number of ramets per clone and the yield of male flowers are positively related to the paternal contribution. The arrangement of the unequal number of plants could be considered in the seed orchard configuration design when establishing the seed orchard, and the reproductive capacity can be used as a predictor of the high yield of parents in the seed orchard.

A higher pollen pollution rate will reduce the genetic gain of offspring. Additional seed orchard management measures are needed to avoid high pollution rates. A first step in avoiding pollen pollution would be to remove the single Chinese fir tree near the seed orchard, implement intensive management of the seed orchard, and set up isolation areas. Although the effective transmission distance of pollen in this study was 5–176 m, the pollen particles were small and light (Ma et al., 2017), and the dispersion distance was greater than 600 m (Chen et al., 1996), which justifies that the isolation area should be greater than 600 m. Although the mating success of Chinese fir had a loose male and female florescence overlap (about 30%), the uneven male parental contribution may lead to the seeds having common ancestors and reduce the stability and adaptability of the progeny stand. Therefore, it is necessary to implement pollen management strategies, such as controlling pollination and supplementing large-scale pollination, to improve the genetic quality of the produced seeds (Kaya et al., 2006; Stoehr et al., 2006; Fernandes et al., 2008; Funda and El-Kassaby, 2012; Chen et al., 2018). Management measures such as delaying flowering can be implemented for parents who do not encounter flowering (Funda and El-Kassaby, 2012). Inbreeding should be avoided in seed orchard production, but inbreeding is inevitable in recurrent selection. This study found that most matings occur within a given distance, and 19.90% occur in the nearest neighbors in the advanced seed orchard of Chinese fir. Moreover, we also found nearly 10% of inbreeding. Therefore, it is requisite to design the layout of seed orchards to separate related parental clones to maximize the spatial distance between them and minimize the impact of inbreeding. Traditional methods, such as adjusted random block design and grouping arrangement, were often used in the layout of Chinese fir seed orchards, but are no longer used in advanced generation seed orchards. Based on the development of computer algorithms, many excellent seed orchard layout designs have emerged, such as COOL (Bell and Fletcher, 1978), MI (Lstibůrek and El-Kassaby, 2010; Lstibůrek et al., 2015), R^2SCR (El-Kassaby et al., 2014), ONA (Chaloupková et al., 2016), and IAPGA (Yang et al., 2020). However, the meeting of male and female gametes at florescence is the premise of mating, and the amount of male and female flowers is the embodiment of fertility. In recent years, new bionic intelligent algorithms have emerged and been optimized, making it possible to design the layout of complex advanced generation seed orchards that combine relationship, florescence, and flower amount.

5 Conclusion

Monitoring genetic diversity and mating patterns is especially important in high-generation seed orchards. Correlation studies of parental clones in seed orchards and pollination dynamics could prove helpful in making management decisions aimed at improving the genetic quality of seeds and in protecting the genetic resources of seed orchards. Our conclusions are as follows. (1) Genetic coancestry among parental clones was detected in the third generation seed orchard of Chinese fir; (2) The seed orchard had a high level of outcrossing (100%), and no selfing offspring were found; (3) The parental contribution in the Chinese fir seed orchard was unbalanced, and the male parental contribution showed a significant positive correlation with the number of ramets of the clone and the number of male flowers of the clone, but was not correlated with the flowering synchronization; (4) The efficient pollination of seed orchards mostly occurs at close range. In addition, we also found a small proportion of mating between related clones. Assessment of pollination dynamics should become a routine procedure in advanced generation seed orchard management to monitor the progress of orchard crop inbreeding and fitness. The results from this research are significant for the management and construction of seed orchards and the formulation of new breeding strategies in the future.

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 in the article.

Author contributions

AD and HW designed the experiments. HW, SZ and XW participated in data collection and phenotype measurement. AD and SZ participated in data analysis and processing. AD and JZ conceived the project and obtained fundings. JZ performed the field management. HW wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the national key research and development project of China of the "14th five-year plan": "Research on Breeding of high-yield, high-quality and high-efficiency new varieties of Chinese Fir".

Acknowledgments

The authors are grateful to the anonymous reviewers and handling Editor for their constructive comments to improve 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.2022.1042290/full#supplementary-material

References

Aronen, T., Nikkanen, T., Harju, A., Tiimonen, H., and Häggman, H. (2002). Pollen competition and seed-siring success in *Picea abies*. *Theor. Appl. Genet.* 104, 638–642. doi: 10.1007/s00122-001-0789-9

Askew, G. R., and Blush, T. D. (1990). Short note: An index of phenological overlap in flowering for clonal conifer seed orchards. *Silvae Genetica* 39, 168–171.

Bell, G. D., and Fletcher, A. M. (1978). Computer organised orchard layouts (COOL) based on the permutated neighbourhood design concept. *Silvae Genet.* 27, 223–225.

Bian, L. M., Huang, D., Zhang, X. F., Tong, X. L., Ye, D. Q., and Shi, J. S. (2020). Analysis on flowering phenology and synchronization indexes of Chinese fir clonal archive. *J. Nanjing Forestry Univ. (Natural Sci. Edition)* 44 (6), 207–212. doi: 10.3969/j.issn.1000-2006.202009016

Chaix, G., Gerber, S., Razafimaharo, V., Vigneron, P., Verhaegen, D., and Hamon, S. (2003). Gene flow estimation with microsatellites in a Malagasy seed orchard of *Eucalyptus grandis*. *Theor. Appl. Genet.* 107, 705–712. doi: 10.1007/s00122-003-1294-0

Chaloupková, K., Stejskal, J., El-Kassaby, Y. A., and Lstibůrek, M. (2016). Optimum neighborhood seed orchard design. *Tree Genet. Genomes* 12, 105. doi: 10.1007/s11295-016-1067-y

Chaloupková, K., Stejskal, J., El-Kassaby, Y., Frampton, J., and Lstibůrek, M. (2019). Current advances in seed orchard layouts: Two case studies in conifers. *Forests* 10 (2), 93. doi: 10.3390/f10020093

Chen, X. Y. (1991). "Observation and analysis of pollen dispersal in Chinese fir seed orchard," in *Symposium on forest genetic improvement* Vol. 19 (Beijing, China: China Forestry Society genetic breeding Society Press).

Chen, X. Y., Li, W. G., Pan, Q. M., and Yang, M. S. (1996). Study on spatial distribution and propagation distance of pollen in Chinese fir seed orchard. *J. Beijing Forestry Univ.* 18 (2), 24–30.

Chen, X., Sun, X., Dong, L., and Zhang, S. (2018). Mating patterns and pollen dispersal in a Japanese larch (*Larix kaempferi*) clonal seed orchard: A case study. *Sci. China Life Sci.* 61, 1011–1023. doi: 10.1007/s11427-018-9305-7

Chen, X., Xu, H., Xiao, F., Sun, S., Lou, Y., Zou, Y., et al. (2021). Genetic diversity and paternity analyses in a 1.5th generation seed orchard of chenshan red-heart Chinese fir. *J. Nanjing Forestry Univ. (Natural Sci. Edition)* 45 (3), 87–92.

Chen, T., Zhang, Z., Chu, X. L., Jin, G. Q., Zhou, Z. C., and Feng, Z. P. (2019). The flowering synchronicity of second-generation clonal seed orchard of masson pine (*Pinus massoniana*). *Scientia Silvae Sinicae* 55 (1), 146–156. doi: 10.11707/j.1001-7488.20190117

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* 1695, 1–9.

Duan, H., Cao, S., Zheng, H., Hu, D., Lin, J., Cui, B., et al. (2017). Genetic characterization of Chinese fir from six provinces in southern China and construction of a core collection. *Sci. Rep.* 7 (1), 13814. doi: 10.1038/s41598-017-13219-0

El-Kassaby, Y. A. (1992). Domestication and genetic diversity - should we be concerned? *Forestry Chronicle* 68, 687–700. doi: 10.5558/tfc68687-6

El-Kassaby, Y. A. (1995). Evaluation of the tree-improvement delivery system: factors affecting genetic potential. *Tree Physiol.* 15, 545–550. doi: 10.1093/treephys/15.7-8.545

El-Kassaby, Y. A. (2000a). "Effect of forest tree domestication on gene pools," in *Forest conservation genetics: Principles and practice* (Canberra, Australia: CSIRO Publishing-CABI Publishing).

El-Kassaby, Y. A. (2000b). Representation of Douglas-fir and western hemlock families in seedling crops as affected by seed biology and nursery crop management practices. *For. Genet.* 7, 305–315.

El-Kassaby, Y. A., Fayed, M., Klápště, J., and Lstibůrek, M. (2014). Randomized, replicated, staggered clonal-row (R^2SCR) seed orchard design. *Tree Genet. Genomes* 10 (3), 555–563. doi: 10.1007/s11295-014-0703-7

El-Kassaby, Y., and Reynolds, S. (1990). Reproductive phenology, parental balance, and supplemental mass pollination in a sitka-spruce seed-orchard. *For. Ecol. Manage.* 31, 45–54. doi: 10.1016/0378-1127(90)90110-W

El-Kassaby, Y. A., and Ritland, Y. A. (1986). The relation of outcrossing and contamination to reproductive phenology and supplemental mass pollination in a Douglas-fir seed orchard. *Silvae Genetica* 35 (5), 240–244.

El-Kassaby, Y. A., and Sziklai, O. (1982). Genetic variation of allozyme and quantitative traits in a selected Douglas-fir (*Pseudotsuga menziesii* var. *menziesii* (Mirb.) Franco) population. *For. Ecol. Manage.* 4, 115–126. doi: 10.1016/0378-1127(82)90009-3

Eriksson, G., Lindgren, D., and Jonsson, A. (1973). Flowering in a clone trial of *Picea abies Karst*. *Stud. For. Suec* 110, 1–45.

Feng, F. J., Sui, X., Chen, M. M., Zhao, D., Han, H. J., and Li, M. H. (2010). Mode of pollen spread in clonal seed orchard of *Pinus koraiensis*. *J. Biophys. Chem.* 01, 33–39. doi: 10.4236/jbpc.2010.11004

Fernandes, L., Rocheta, M., Cordeiro, J., Pereira, S., Gerber, S., Oliveira, M. M., et al. (2008). Genetic variation, mating patterns and gene flow in a *Pinus pinaster* aiton clonal seed orchard. *Ann. For. Sci.* 65, 706–706. doi: 10.1051/forest:2008049

Funda, T., Chen, C. C., Cherdsak, L., M.A.Kenawy, A., and El-Kassaby, Y. A. (2008). Pedigree and mating system analyses in a western larch (*Larix occidentalis* nutt.) experimental population. *Ann. For. Sci.* 65, 705. doi: 10.1051/forest2008055

Funda, T., and El-Kassaby, Y. A. (2012). Seed orchard genetics. *CAB Reviews: Perspect. Agriculture Veterinary Science Nutr. Natural Resour.* 7 (13), 1–23. doi: 10.1079/PAVSNNR20127013

Funda, T., Lstiburek, M., Lachout, P., Klapste, J., and El-Kassaby, Y. A. (2009). Optimization of combined genetic gain and diversity for collection and deployment of seed orchard crops. *Tree Genet. Genomes* 5, 583–593. doi: 10.1007/s11295-009-0211-3

Funda, T., Wennström, U., Almqvist, C., Andersson Gull, B., and Wang, X.-R. (2016). Mating dynamics of scots pine in isolation tents. *Tree Genet. Genomes* 12, 112. doi: 10.1007/s11295-016-1074-z

Funda, T., Wennström, U., Almqvist, C., Torimaru, T., Gull, B. A., and Wang, X.-R. (2015). Low rates of pollen contamination in a scots pine seed orchard in Sweden: the exception or the norm? *Scandinavian J. ofForest Res.* 30 (7), 573–586. doi: 10.1080/02827581.2015.1036306

González-Martínez, S., Salvador, L., Agúndez, D., Alia, R., and Gil, L. (2001). *Geographical variation of gene diversity of pinus pinaster ait. in the Iberian peninsula* (Dordrecht-Boston-London: Kluwer Academic Publishers).

Grattapaglia, D., Do Amaral Diener, P. S., and Dos Santos, G. A. (2014). Performance of microsatellites for parentage assignment following mass controlled pollination in a clonal seed orchard of loblolly pine (*Pinus taeda* l.). *Tree Genet. Genomes* 10, 1631–1643. doi: 10.1007/s11295-014-0784-3

He, G., Qi, M., Cheng, Y., Xu, Z., Luo, X., and He, B. (2016). Methods for parent apolegamy of Chinese fir in cross breeding. *Acta Agriculturae Universitatis Jiangxiensis* 38 (4), 646–535. doi: 10.13836/j.jjau.2016092

Heuchel, A., Hall, D., Zhao, W., Gao, J., Wennström, U., and Wang, X.-R. (2022). Genetic diversity and background pollen contamination in Norway spruce and scots pine seed orchard crops. *Forestry Res.* 2, 1–12. doi: 10.48130/FR-2022-0008

Hodge, G. R., and White, T. L. (1993). Advanced-generation wind-pollinated seed orchard design. *New Forests* 7, 213–236. doi: 10.1007/BF00127387

Huang, L. S., Song, J. Y., Sun, Y. Q., Gao, Q., Jiao, S. Q., Zhou, S. S., et al. (2018). Pollination dynamics in a *Platycladus orientalis* seed orchard as revealed by partial pedigree reconstruction. *Can. J. For. Res.* 48 (8), 952–957. doi: 10.1139/cjfr-2018-0077

Jiang, S., Shi, J. S., Zhang, Y. H., Zhou, C. G., and Hu, C. Y. (1986). "A preliminary study on pollen transmission distance of Cunninghamia lanceolata (Lamb.) hook," in *The fifth symposium on forest tree genetics and breeding in China*, vol. 94. (Harbin, China: Northeast Forestry University Press).

Kalinowski, S., Taper, M., and Marshall, T. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294X.2007.03089.x

Kaya, N., Isik, K., and Adams, W. T. (2006). Mating system and pollen contamination in a *Pinus brutia* seed orchard. *New Forests* 31, 409–416. doi: 10.1007/s11056-005-0876-x

Kess, T., and El-Kassaby, Y. A. (2015). Estimates of pollen contamination and selfing in a coastal Douglas-fir seed orchard. *Scandinavian J. For. Res.* 30 (4), 266–275. doi: 10.1080/02827581.2015.1012112

Korecký, J., and El-Kassaby, Y. A. (2016). Pollination dynamics variation in a Douglas-fir seed orchard as revealed by microsatellite analysis. *Silva Fennica* 50 (4), 1682. doi: 10.14214/sf.1682

Laucou, V., Launay, A., Bacilieri, R., Lacombe, T., Adam-Blondon, A. F., Berard, A., et al. (2018). Extended diversity analysis of cultivated grapevine vitis vinifera with 10K genome-wide SNPs. *PloS One* 13, e0192540. doi: 10.1371/journal.pone.0192540

Liesebach, H., Liepe, K., and Bäucker, C. (2021). Towards new seed orchard designs in Germany – a review. *Silvae Genetica* 70, 84–98. doi: 10.2478/sg-2021-0007

Lindgren, D., Danusevic˘ius, D., and Rosvall, O. (2009). Unequal deployment of clones to seed orchards by considering genetic gain, relatedness and gene diversity. *Forestry* 82 (1), 17–28. doi: 10.1093/forestry/cpn033

Lindgren, D., and Prescher, F. (2005). Optimal clone number for seed orchards with tested clones. *Silvae Genetica* 54, 80–92. doi: 10.1515/sg-2005-0013

Li, W., Wang, X., and Li, Y. (2011). Stability in and correlation between factors influencing genetic quality of seed lots in seed orchard of *Pinus tabuliformis* carr. over a 12-year span. *PloS One* 6, e23544. doi: 10.1371/journal.pone.0023544

Li, Y. X., Wang, Z. S., Sui, J. K., Zeng, Y. F., Duan, A. G., and Zhang, J. G. (2015). Isolation and characterization of microsatellite loci for *Cunninghamia lanceolata* (Lamb.) hook. *Genet. Mol. Res.* 14, 453–456. doi: 10.4238/2015.January.23.19

Li, X., Wang, L. B., Wen, Y. F., Lin, J., Wu, X. T., Yuan, M. L., et al. (2020). Genetic diversity of of Chinese fir (*Cunninghamia lanceolata*) breeding population among different generations. *Scientia Silvae Sinicae* 56 (11), 53–61. doi: 10.11707/j.1001-7488.20201106

Lstibůrek, M., and El-Kassaby, Y. (2010). Minimum-inbreeding seed orchard design. *For. Sci.* 56, 603–608.

Lstibůrek, M., Stejskal, J., Misevicius, A., Korecky, J., and El-Kassaby, Y. (2015). Expansion of the minimum-inbreeding seed orchard design to operational scale. *Tree Genet. Genomes* 11, 12. doi: 10.1007/s11295-015-0842-5

Ma, K. B., Tang, L., and Cui, J. W. (2017). Change of morphological structure during pollen development process in *Cunninghamia lanceolate*. *Guihaia* 37 (10), 1342–1347. doi: 10.11931 /guihaia.gxzw201612042

McKeand, S. E., and Bridgwater, F. E. (1998). A strategy for the third breeding cycle of loblolly pine in the southeastern U.S. *Silvae Genetica.* 47 (4), 223–234.

Moriguchi, Y., Taira, H., Tani, N., and Tsumura, Y. (2004). Variation of paternal contribution in a seed orchard of *Cryptomeria japonica* determined using microsatellite markers. *Can. J. For. Res.* 34, 1683–1690. doi: 10.1139/x04-029

Moriguchi, Y., Tani, N., Itoo, S., Kanehira, F., Tanaka, K., Yomogida, H., et al. (2005). Gene flow and mating system in five *Cryptomeria japonica* d. don seed orchards as revealed by analysis of microsatellite markers. *Tree Genet.* 1 (4), 174–183. doi: 10.1007/s11295-005-0023-z

Mullin, T. J., Persson, T., Abrahamsson, S., and Andersson Gull, B. (2019). Effects of inbreeding depression on seed production in scots pine (*Pinus sylvestris*). *Can. J. For. Res.* 49, 854–860. doi: 10.1139/cjfr-2019-0049

Muñoz-Gutiérrez, L., Vargas-Hernández, J. J., López-Upton, J., Ramírez-Herrera, C., and Jiménez-Casas, M. (2020). Clonal variation in phenological synchronization and cone production in a *Pinus patula* seed orchard. *Silvae Genetica* 69, 130–138. doi: 10.2478/sg-2020-0018

Nielsen, U. B., and Hansen, O. K. (2011). Genetic worth and diversity across 18 years in a nordmann fir clonal seed orchard. *Ann. For. Sci.* 69, 69–80. doi: 10.1007/s13595-011-0159-y

Nikkanen, T., Aronen, T., Häggman, H., and Venäläinen, M. (2000). Variation in pollen viability among *Picea abies* genotypes – potential for unequal paternal success. *Theor. Appl. Genet.* 101 (4), 511–518. doi: 10.1007/s001220051510

Nikkanen, T., and Ruotsalainen, S. (2000). Variation in flowering abundance and its impact on the genetic diversity of the seed crop in a Norway spruce seed orchard. *Silva Fennica* 34 (3), 626. doi: 10.14214/sf.626

Ou, J. L. (2012). *Pollen dispersal, in vitro competition and open-pollinated offspring fitness in natural populations of betula alnoides*. (Chinese Academy of Forestry: Master).

Pakull, B., Eusemann, P., Wojacki, J., Ahnert, D., and Liesebach, H. (2021). Genetic diversity of seeds from four German Douglas fir (*Pseudotsuga menziesii*) seed orchards. *Eur. J. For. Res.* 140, 1543–1557. doi: 10.1007/s10342-021-01419-3

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in excel. population genetic software for teaching and research–an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Ruņģis, D., Luguza, S., Bāders, E., Šķipars, V., and Jansons, Ā. (2019). Comparison of genetic diversity in naturally regenerated Norway spruce stands and seed orchard progeny trials. *Forests* 10 (10), 926. doi: 10.3390/f10100926

Sønstebø, J. H., Tollefsrud, M. M., Myking, T., Steffenrem, A., Nilsen, A. E., Edvardsen, Ø.M., et al. (2018). Genetic diversity of Norway spruce (*Picea abies* (L.) karst.) seed orchard crops: Effects of number of parents, seed year, and pollen contamination. *For. Ecol. Manage.* 411, 132–141. doi: 10.1016/j.foreco.2018.01.009

Sheng, W. T. (2014). *Plantation forest and their silviculture systems in China* (Beijing, China: China Forestry Press).

Shi, J. S. (1994). Present situation of genetic improvement of *Cunninghamia lanceolata* in fujian province and its technical countermeasures. *Fujian Forestry Sci. Technol.* 21 (3), 28–31.

Sk Lai, B., Funda, T., Liewlaksaneeyanawin, C., Klápště, J., Niejenhuis, A., Cook, C., et al. (2010). Pollination dynamics in a Douglas-fir seed orchard as revealed by pedigree reconstruction. *Ann. For. Sci.* 67, 808–808. doi: 10.1051/forest/2010044

Song, J., Ratcliffe, B., Kess, T., Lai, B. S., Korecky, J., and El-Kassaby, Y. A. (2018). Temporal quantification of mating system parameters in a coastal Douglas-fir seed orchard under manipulated pollination environment. *Sci. Rep.* 8, 11593. doi: 10.1038/s41598-018-30041-4

Sork, V. L., Davis, F. W., Smouse, P. E., Apsit, V. J., Dyer, R. J., Fernandez, J. F., et al. (2002). Pollen movement in declining populations of California valley oak, quercus lobata: where have all the fathers gone? *Mol. Ecol.* 11, 1657–1668. doi: 10.1046/j.1365-294X.2002.01574.x

Stoehr, M., Mehl, H., Nicholson, G., Pieper, G., and Newton, C. (2006). Evaluating supplemental mass pollination efficacy in a lodgepole pine orchard in British Columbia using chloroplast DNA markers. *New Forests* 31, 83–90. doi: 10.1007/s11056-004-5398-4

Stoehr, M. U., Orvar, B. L., Vo, T. M., Gawley, J. R., Webber, J. E., and Newton, C. H. (1998). Application of a chloroplast DNA marker in seed orchard management evaluations of Douglas-fir. *Can. J. For. Res.* 28, 187–195. doi: 10.1139/x97-201

Stoehr, M., Ott, P., and Woods, J. (2014). Inbreeding in mid-rotation coastal Douglas-fir: implications for breeding. *Ann. For. Sci.* 72, 195–204. doi: 10.1007/s13595-014-0414-0

Torimaru, T., Wennstrom, U., Lindgren, D., and Wang, X. R. (2012). Effects of male fecundity, interindividual distance and anisotropic pollen dispersal on mating success in a scots pine (*Pinus sylvestris*) seed orchard. *Heredity (Edinb)* 108, 312–321. doi: 10.1038/hdy.2011.76

Wang, J. (2011). COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Mol. Ecol. Resour.* 11, 141–145. doi: 10.1111/j.1755-0998.2010.02885.x

Wang, T., Aitken, S. N., Woods, J. H., Polsson, K., and Magnussen, S. (2004). Effects of inbreeding on coastal Douglas fir growth and yield in operational plantations: a model-based approach. *Theor. Appl. Genet.* 108, 1162–1171. doi: 10.1007/s00122-003-1534-3

Waples, R. S., and Do, C. (2008). LDNE: A program for estimating effective population size from data on linkage disequilibrium. *Mol. Ecol. Resour.* 8, 753–756. doi: 10.1111/j.1755-0998.2007.02061.x

Wei, Z. G., Qu, Z. S., Hou, Y. Y., Zhang, L. J., Yang, C. P., and Wei, H. R. (2015). Genetic diversity and paternal analysis of openpollinated progenies of *Larix olgensis* seed orchard. *J. Nat. Sci.* 1, e19.

Wen, Y. F., Han, W. J., Zhou, H., and Xu, G. B. (2015). SSR mining and development of EST-SSR markers for *Cunninghamia lanceolata* based on transcriptome sequences. *Scientia Silvae Sinicae* 51 (11), 41–49. doi: 10.11707/j.1001-7488.20151106

Wheeler, N., and Jech, K. (1992). The use of electrophoretic markers in seed orchard research. *New Forests* 6, 311–328. doi: 10.1007/BF00120650

Woods, J. H., and Heaman, J. C. (1989). Effect of different inbreeding levels on filled seed production in Douglas-fir. *Can. J. For. Res.* 19, 54–59. doi: 10.1139/x89-007

Woods, J. H., Wang, T., and Aitken, S. (2001). Effects of inbreeding on coastal Douglas fir nursery performance. *Silvae Genetica* 51, 163–170.

Xu, Q., Wu, C., and Xu, Z. (2015). Analysis of Chinese fir full diallel progeny genetic variation. *Hunan Forestry Sci. Technol.* 42 (6), 7–12. doi: 10.3969/j.issn.1003-5710.2015.06.002

Xu, Q. Q., Xu, Z. K., Zhang, X., Rong, J. P., and Gu, Y. C. (2013). Study and formulation of the technical regulations for construction of the Chinese fir seed orchard. *Hubei forestry Sci. Technol.* 42 (4), 84–87.

Yang, B., Niu, S., El-Kassaby, Y. A., and Li, W. (2021). Monitoring genetic diversity across *Pinus tabuliformis* seed orchard generations using SSR markers. *Can. J. For. Res.* 51, 1534–1540. doi: 10.1139/cjfr-2020-0479

Yang, B., Sun, H., Qi, J., Niu, S., El-Kassaby, Y. A., and Li, W. (2020). Improved genetic distance-based spatial deployment can effectively minimize inbreeding in seed orchard. *For. Ecosyst.* 7, 10. doi: 10.1186/s40663-020-0220-0

Yang, H. B., Zhang, R., and Zhou, Z. C. (2016). Genetic diversity and mating system in a seed orchard of *Schima superba*. *Scientia Silvae Sinicae* 52 (12), 66–73.

Yang, H., Zhang, R., and Zhou, Z. (2017). Pollen dispersal, mating patterns and pollen contamination in an insect-pollinated seed orchard of *Schima superba* gardn. et champ. *New Forests* 48, 431–444. doi: 10.1007/s11056-017-9568-6

Zhang, Z. W. (2005). *Studies on Chinese fir reproduction characteristics* (Huazhong Agricultural University: Doctor).

Zhang, X., Tang, X. J., and Cheng, G. Y. (2016). Flowering synchronization of *Phellodendrom amurense* in seed orchard. *J. Northeast Forestry Univ.* 44 (7), 46–50. doi: 10.13759/j.cnki.dlxb.2016.07.010

Zheng, H., Hu, D., Wei, R., Yan, S., and Wang, R. (2019). Chinese Fir breeding in the high-throughput sequencing era: Insights from SNPs. *Forests* 10 (8), 681. doi: 10.3390/f10080681

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fpls.2022.1068883

OPEN ACCESS

EDITED BY
Satoshi Watanabe,
Saga University, Japan

REVIEWED BY
Narayanan Manikanda Boopathi,
Tamil Nadu Agricultural University, India
Lalit Arya,
National Bureau of Plant Genetic Resources (ICAR), India
Lohithaswa Hirenallur Chandappa,
University of Agricultural Sciences, Bangalore, India

*CORRESPONDENCE
Nepolean Thirunavukkarasu
tnepolean@gmail.com

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

RECEIVED 13 October 2022
ACCEPTED 15 December 2022
PUBLISHED 10 January 2023

CITATION
Semalaiyappan J, Selvanayagam S, Rathore A, Gupta SK, Chakraborty A, Gujjula KR, Haktan S, Viswanath A, Malipatil R, Shah P, Govindaraj M, Ignacio JC, Reddy S, Singh AK and Thirunavukkarasu N (2023) Development of a new AgriSeq 4K mid-density SNP genotyping panel and its utility in pearl millet breeding. *Front. Plant Sci.* 13:1068883. doi: 10.3389/fpls.2022.1068883

Development of a new AgriSeq 4K mid-density SNP genotyping panel and its utility in pearl millet breeding

Janani Semalaiyappan[1], Sivasubramani Selvanayagam[2], Abhishek Rathore[3], SK Gupta[2], Animikha Chakraborty[1], Krishna Reddy Gujjula[4], Suren Haktan[4], Aswini Viswanath[1], Renuka Malipatil[1], Priya Shah[1], Mahalingam Govindaraj[5], John Carlos Ignacio[6], Sanjana Reddy[1], Ashok Kumar Singh[7] and Nepolean Thirunavukkarasu[1]*

[1]Genomics and Molecular Breeding Lab, ICAR-Indian Institute of Millets Research, Rajendranagar, India, [2]Accelerated Crop Improvement, International Crop Research Institute for the Semi-Arid Tropics (ICRISAT), Hyderabad, India, [3]Excellence in Breeding (EiB) Platform, The International Maize and Wheat Improvement Center (CIMMYT), El Batán, Mexico, [4]Bioinformatics, Thermo Fisher Scientific, Austin, TX, United States, [5]HarvestPlus, International Center for Tropical Agriculture, Cali, Colombia, [6]Department of Horticulture and Crop Science, The Ohio State University, Wooster, OH, United States, [7]ICAR-Indian Agricultural Research Institute, New Delhi, India

Pearl millet is a crucial nutrient-rich staple food in Asia and Africa and adapted to the climate of semi-arid topics. Since the genomic resources in pearl millet are very limited, we have developed a brand-new mid-density 4K SNP panel and demonstrated its utility in genetic studies. A set of 4K SNPs were mined from 925 whole-genome sequences through a comprehensive in-silico pipeline. Three hundred and seventy-three genetically diverse pearl millet inbreds were genotyped using the newly-developed 4K SNPs through the AgriSeq Targeted Genotyping by Sequencing technology. The 4K SNPs were uniformly distributed across the pearl millet genome and showed considerable polymorphism information content (0.23), genetic diversity (0.29), expected heterozygosity (0.29), and observed heterozygosity (0.03). The SNP panel successfully differentiated the accessions into two major groups, namely B and R lines, through genetic diversity, PCA, and structure models as per their pedigree. The linkage disequilibrium (LD) analysis showed Chr3 had higher LD regions while Chr1 and Chr2 had more low LD regions. The genetic divergence between the B- and R-line populations was 13%, and within the sub-population variability was 87%. In this experiment, we have mined 4K SNPs and optimized the genotyping protocol through AgriSeq technology for routine use, which is

cost-effective, fast, and highly reproducible. The newly developed 4K mid-density SNP panel will be useful in genomics and molecular breeding experiments such as assessing the genetic diversity, trait mapping, backcross breeding, and genomic selection in pearl millet.

KEYWORDS

pearl millet, mid-density SNP, AgriSeq technology, high-throughput genotyping, genomics

Introduction

Pearl millet (*Pennisetum glaucum* (L) R. Br., syn. *Cenchrus americanus* (L.) Morrone) is a strategic climate-resilient C4 crop. It has an inherent ability to provide sustainable yield even in harsh ecologies, making it an economically secure and favorable crop for farmers in semi-arid and arid regions of the world. Pearl millet is also known for nutritional security as it is competent to address malnutrition issues (Nambiar et al., 2011; Kanatti et al., 2014). In traditional plant breeding, superior genotypes have been selected visually. With the discovery of genetic markers, crop breeding heavily depends upon the reliable and cost-effective marker system. Developing viable markers and genotyping platforms in any crop is imperative to accelerate the varietal turnover.

Various marker genotyping methods such as expressed sequence tags- derived simple sequence repeats (EST-SSRs) (Senthilvel et al., 2008; Rajaram et al., 2013), genomic simple sequence repeats (gSSRs), (Qi et al., 2004), DArT array Technology (DArTs) (Senthilvel et al., 2010; Supriya et al., 2011), and single nucleotide polymorphisms (SNPs) (Sehgal et al., 2012) have been developed and used to characterize the pearl millet genome, identification of quantitative trait loci (QTLs) and marker-assisted breeding (MAB)

activities. The EST-SSRs were developed and utilized in the genetic mapping and MAB programs targeting the traits such as yield and drought resistance in pearl millet (Senthilvel et al., 2010). Through DArT and SSRs marker systems, the QTLs for grain iron and zinc content and rust resistance were identified in the RILs population of pearl millet (Ambawat et al., 2016; Kumar et al., 2016; Kumar et al., 2018). Later, the EST-derived SNPs markers were developed in pearl millet and deployed for identifying the major QTLs-associated candidate genes for drought tolerance using two mapping populations (Sehgal et al., 2012). The *de-novo* sequencing of the pearl millet genome (1.79 Gb), (Varshney et al., 2017) provided an opportunity to explore the genome comprehensively and develop new genomic resources. They have a great potential for understanding the genetic architecture and quantitative traits as well as improving such traits in pearl millet.

There are mainly three types of throughput platforms, namely high-density (10's of thousands of SNPs), mid-density (a few thousand SNPs), and low-density (less than 100 SNPs), generally used for genotyping purposes. The high-density platforms are usually applicable for whole genome studies and trait mapping but are very tedious, expensive, laborious, and time-consuming. High-density platforms such as Illumina, PacBio, genotyping-by-sequencing (GBS), and restriction site-associated DNA sequencing (RAD) provide tens of thousands of genome-wide SNPs. The mid-density platforms, which include AgriSeq (Koelewijn, 2018), DartTag (Kilian et al., 2012; Ren et al., 2015), and RiCA (Arbelaez et al., 2019) have their applicability in genotyping 1000 to 5000 SNPs and are highly advantageous in terms of being time-efficient and user-friendly with rapid data interpretation. Low-density platforms are mainly used to track specific QTLs or genes. TaqMan and KASP™ (Kompetitive Allele-Specific Polymerase chain reaction) are the widely used low-density SNP platforms (Thomson et al., 2012; Ganal et al., 2019). Array-based SNP chips are developed in several crops by including the identified SNPs printed on the chip. For example, in maize, Illumina developed a golden gate assay Illumina® 1536 SNP chip and Illumina® MaizeSNP50 Beadchip (www.illumina.com/maizeSNP50, Wu et al., 2014) and has been used for various genetic applications (Thirunavukkarasu et al., 2013; Nepolean et al., 2014; Thirunavukkarasu et al., 2017).

The mid-density genotyping approach is highly efficient, informative, and cost-effective as it can be used in various genomics and molecular breeding experiments such as diversity assessment, trait mapping, marker-assisted breeding, and genomic selection. Although a crop with immense economic and social importance, pearl millet has been neglected for a long time, and not enough efforts have been made to explore genomic resources. Hence, the objectives of the experiment were to mine a mid-density panel of 4000 SNPs from 925 whole genome sequences of pearl millet, to develop a functional, robust, and reproducible genotyping protocol through AgriSeq technology, to characterize the newly developed SNPs in a set of 373 genetically diverse pearl millet B and R lines and to demonstrate its utility in genetic studies.

Material and methods

Mining SNPs

The whole-genome resequences (WGRS) from two sets of pearl millet accessions (Varshney et al., 2017) were considered for developing the mdi-density marker panel– 1. The Pearl Millet Inbred Germplasm Association Panel (PMiGAP) representing 345 genotypes (263 landraces or traditional cultivars, 46 breeding lines, 25 advanced or improved cultivars, and 11 accessions with unknown biological status) (hereafter "Group A" genotypes) and 2. Diverse breeding lines representing 580 genotypes (260 B and 320 R lines) (hereafter "Group B" genotypes), assuming to accommodate possible racial and geographical representation of pearl millet breeding diversity.

Extraction of SNPs from the Group A and B genotypes

The variant data (32 million SNPs) from the Groups A and B genotypes were retrieved from the pearl millet genome project (Varshney et al., 2017). These variants were pruned for site coverage (90%) and minimum minor allele frequency (0.01) using Tassel (version 5.2.51), which resulted in 276K bi-allelic SNP markers. Filtration of the markers mentioned above for specificity, polymorphic information content (PIC), flanking SNPs on windows (50bp), and presence of flanking SSRs (with pearl millet reference genome v1.1) resulted in 67K SNPs. For the specificity check, we used Bowtie (v1.1.2) with no variations allowed upon mapping. Only the SNPs falling on exons were taken forward from the filtered variants. This marker set was further put on two independent random selection exercises, using the purity tool (https://bitbucket.org/jcignacio/purity/wiki/Home), for picking an initial set of 6000 markers each that can carry maximum genetic information. These two sets (from different iterations) filtered for high LD (LDBlockShow) and redundant markers, resulting in 2000 SNPs each for Group A (2K set I) and B genotypes (2K set II), respectively. The 2K sets I and II together formed a 4K mid-density SNP panel (4K set). The 4K panel were functionally annotated using SnpEff (v4.3t) with predicted pearl millet reference annotations (Figure 1).

SNP assay design

The selected SNPs were passed through AgriSeq's design quality control process. The quality check was performed using the pearl millet reference genome (accession: GCA_002174835.2) and then submitted to the primer design phase. The primer designs were *in-silico* checked for specificity and sensitivity of the intended target/marker regions using pearl millet reference genome. Finally, 4000 SNPs were selected to constitute the custom 4K SNP GBS pearl millet panel.

Sequencing

The AgriSeq targeted-GBS solution utilizes a highly efficient multiplexed PCR chemistry where hundreds to thousands of markers can be targeted and uniformly amplified in a single reaction. Three eighty-four samples were prepared for sequencing using the AgriSeq HTS Library Kit (A34143-Life Technologies). In short, DNA concentrations were normalized to 3.3 ng/μL for a total of 10 ng DNA per 10 μL reaction. Normalized DNA was combined with the AgriSeq custom primer panel and AgriSeq amplification master mix. For amplification of genomic targets, the following thermocycling programs were used; 99°C for 2 minutes, then 15 cycles of 99°C for 15s and 60°C for 4 minutes. Amplicons were prepared for ligation with pre-ligation enzyme digestion at 50°C for 10 minutes, 55°C for 10 minutes, and 60°C for 20 minutes. IonCode™ Barcode Adapters 385-768 Kit (A36546-Life Technologies) were ligated to the digested products with barcoding enzyme and buffer. Labeled amplicons were then pooled, cleaned up, amplified, and normalized. Following library preparation, libraries were loaded onto an Ion 540™ sequencing Chip Kit (A42849) *via* the Ion 540™ Kit-Chef (A43541-Life Technologies) and Ion Chef. Sequencing was

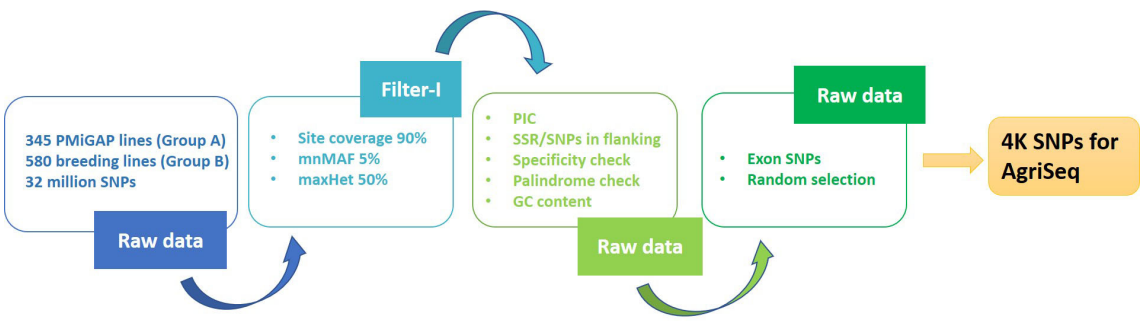


FIGURE 1
In-silico pipeline used in extraction of 4K mid-density SNPs from the 925 genetically diverse pearl millet genome sequences.

performed on the Ion S5 system (Thermo Fisher, Inc. Waltham, MA). After sequencing, genotyping was performed automatically by Torrent Variant Caller (TVC) on the Torrent Suite Server (TS).

Genotyping

The variant calling pipeline is fully automated and optimized for analyzing Ion Torrent sequencing data on Torrent Suite Server (Thermo Fisher Scientific). This workflow comprises several series of steps. First, signal processing files are automatically transferred from the sequencing platform to the S5 server and then converted to raw reads (FASTQ). After that, the sequenced reads were de-multiplexed to individual samples using the barcode sequences. For each sample, the sequenced reads from the targeted regions were mapped to the pearl millet reference genome using TMAP- Torrent Mapping Alignment Program (https://github.com/iontorrent/TS/tree/master/Analysis/TMAP) followed by genotyping using TVC-Torrent Variant Caller (https://github.com/iontorrent/Torrent-Variant-Caller-stable). The genotypes were reported in different formats TOP, TOP/BOT, and actual alleles using AgriSum Toolkit, an AgriSeq TS plugin (https://assets.thermofisher.com/TFS-assets/LSG/manuals/MAN0018917_AgriSum_plugin_UB.pdf).

Data analysis

SNP statistics

A set of 373 pearl millet inbreds consisting of 195 B-lines and 182 R-lines received from ICRISAT, Hyderabad (Supplementary Table S1) were subjected to genotyping using the newly developed 4K SNP panel. The genotyping data were analyzed for several metrics, namely PIC, Nei's genetic diversity (GD), minor allele frequency (MAF), expected heterozygosity (He), and observed heterozygosity (Ho). The parameters mentioned above were calculated using the SnpReady package in R (Granato et al., 2018).

Principal component analysis

PCA was conducted using the snpgdsPCA function available in SNPRelate (Zheng et al., 2012). The percentage of variation was calculated for the first 15 principal components, and the genotypes were plotted on a three-dimensional scale using the first three components.

Analysis of molecular variance

The variance at the molecular level of 373 genotypes between and within B- and R-line groups was analyzed through GENEALEX version 6.503 (Peakall and Smouse, 2006) with 999 permutations of the data set using PhiPT value (an analog of fixation index FST). PhiPT in AMOVA, a measure that provides significant insights into the evolutionary processes that influence the structure of genetic variation within and among subgroups, was used to calculate the degree of genetic divergence. PhiPT value represents the ratio of the variance within subgroups to the overall variance between subgroups (Keneni et al., 2012). The high PhiPT value indicates the more significant differences between the subgroups.

Diversity assessment

DARwin (version 6.0.9) (Perrier and Jacquemoud-Collet, 2006) was used for measuring the genetic diversity among the accessions. The unweighted neighbor-joining approach was used to visualize the phylogenetic tree from the dissimilarity coefficient based on a simple matching approach.

Linkage disequilibrium

The extent of the LD in the 373 genotypes in all three set SNP markers was evaluated using TASSEL 5 (Bradbury et al., 2007). For each pair of SNP markers, the squared correlation coefficient (r^2), which measures the correlation between alleles at two loci, was computed along with its corresponding P-value. The SNPs with all MAF, 15% heterozygotes, and 20% missing were included in this analysis. LD values of all pair-wise SNPs were shown in triangle LD plots using TASSEL the genome-wide and chromosome-wise LD patterns.

Population structure

The population structure of the accessions was estimated using an MCMC (Markov Chain Monte Carlo) model implemented in STRUCTURE version 2.3.4 (Pritchard et al., 2000). The data set was evaluated for each K value (2 to 10) with five iterations. The burn-in and MCMC replication numbers were set to 200000 for each run. The most probable K value was determined using the log probability of the data [LnP(D)] and delta K(ΔK) in Structure Harvester (Earl and Vonholdt, 2012). After the optimum K was determined, the graphical representation of the population structure was displayed using the CLUMPACK beta version (Kopelman et al., 2015).

Results

Basic statistics

For the preliminary statistics, the genotypic data from the two SNP sets, 2K set I and 2K set II, were compared since they were derived from two different sets of genotypes. Then, the combined genotypic data, the 4K set, was used for the remaining statistical studies. The data were used to calculate the marker call rate, i.e., the percentage of samples for a particular marker that generate a genotype call. The mean and the median marker call rate for the 2K set I was 93.8% and 98.9%, respectively, while it was 82.9% and 98.9%, respectively, for the 2K set II. For the combined 4K Set, the mean and the median marker call rate was 88.2% and 98.1%, respectively.

In 2K sets I and II, most SNPs were located on chromosomes 2, 1, and 3. Chr2 had comparatively more SNPs in both sets (set I-17.9% and set II-18.8%) (Figure 2). The major allele frequency ranged from 0.5 to 0.99 in both sets, while set I had a higher mean frequency (0.86) over set II (0.73). On the other hand, higher MAF was observed in set II (0.2) over set I (0.14). In set I, 815 detected SNPs was involved in various biological processes contributing to 40.8% of total SNPs, while 31.3% (625) and 43.4% (868) of SNPs were involved in molecular function and cellular components, respectively. Set II showed a similar number of SNPs (832, 39.39%) involved in various biological processes. SNPs involved in molecular function were higher (1137, 53.83%), and cellular components were lesser (344, 16.28%) when compared to Set I (Figure 2). The SNPs identified in the downstream region were 645 and 685, upstream regions were 482 and 651, exon regions were 237 and 293, and intron regions were 636 and 483 for sets I and II, respectively.

The PIC, Ho, and He values were determined using the SNPReady for the data 2K set I, 2K set II, and combined 4K set (Table 1). The PIC for both the sets, 2K set I and II, ranged from 0 to 0.38, while the mean PIC was high in 2K set II (0.3). When comparing sets I and II, 62% of the SNPs showed more than average PIC in 2K set II. About 58% from set II and 44% of SNPs from the 4K set possessed a PIC value that was more than the average. R lines of 2K set II showed the highest mean PIC of 0.28 among all sets, while B-lines showed the highest PIC value (0.27) in the same data set.

The observed heterozygosity (Ho) in 2K set I and 2K set II and 4K set ranged from 0 to 0.21, 0 to 0.41 and 0 to 0.31, while the expected heterozygosity (He) ranged from 0 to 0.21, 0 to 0.38 and 0to 0.3, respectively. 2K set II showed the highest average observed and expected heterozygosity (Ho= 0.04 and He= 0.38). When comparing the B- and R-lines groups, the B-lines had more average Ho in 2K set II and 4K set than the R-lines, whereas the R lines had more average He over the B-lines among all three data sets (Table 1).

Genetic diversity, PCA, and population structure analysis

The grouping behavior of B and R lines was characterized using the 4K set data through genetic diversity, principal component analysis, and structure models.

The SNP frequency-based genetic dissimilarity matrix available in DARwin-6.0 was employed to study the genetic diversity among the 373 genotypes, which included 195 B- and 182-R lines. The NJ-based statistics grouped all the B- and R-lines into two clear-cut major groups (Figure 3) with additional sub-clusters in the respective groups. The first major group (G-I) consisted of 191 B- and 21 R-lines and the second major group (G-II) consisted of 155 R- and 4 B-lines. We also found that one set of R-lines (6 genotypes) formed a small third group. The G-I was further separated into three sub-groups (SG-IA, SG-IA, and SG-IC). SG-IA had 167 B- and 21 R-lines, while SG-IB and SG-IC had 11 and 13 B-lines, respectively. The second major group G-II further grouped the R lines into three subgroups. SG-IIA had 147 R-lines and one B-line, SG-IIB consisted of 2 R-lines and 3 B-lines, and SG-IIC formed by 6 R-lines.

We have also noticed cross-grouping of genotypes, where the R-lines, namely, ICMR 100258, ICMR 102545, ICMR 102497, ICMR 102499, ICMR 11555, ICMR 14111, ICMR 16444, ICMR 16333, ICMR 19888, ICMR 15111, ICMR 07333, ICMR 0755, ICMR 10111, ICMR 13666, ICMR 11888, ICMR

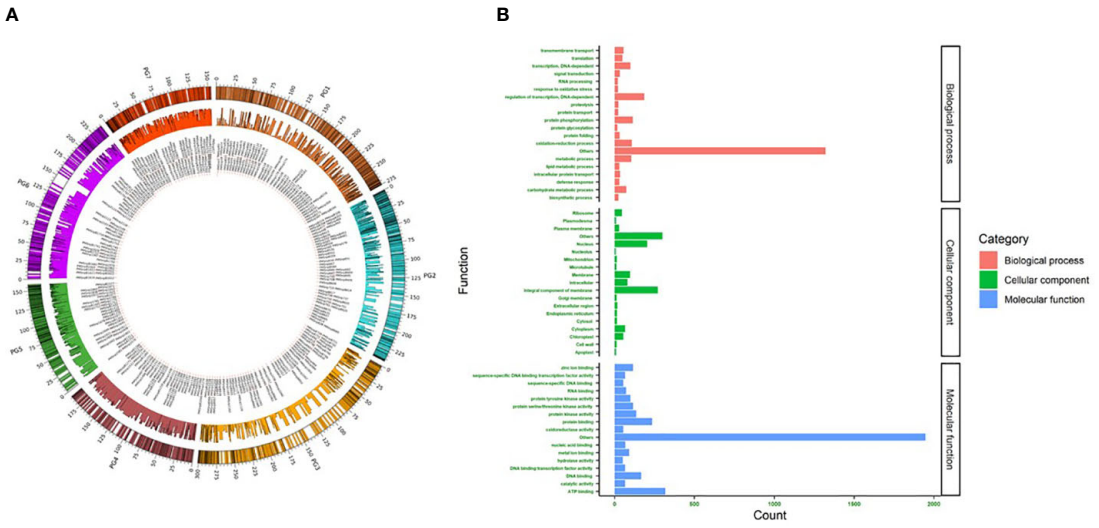


FIGURE 2
(A) Circos plot showing the location and uniform distribution of 4K mid-density SNPs (outer circle), SNP IDs (inner most circle), and PIC values (middle circle) across pearl millet chromosomes. **(B)** Gene ontology of 4K SNPs classified as biological process, cellular component and molecular function.

TABLE 1 Characteristics of the 2K Set I, 2K Set II and 4K Set of SNPs genotyped in a set of 373 B- and R-lines of pearl millet.

	Description	2K Set I	2K Set II	4K Set
PIC	Overall range	0 - 0.38	0 - 0.38	0 - 0.38
	Mean	0.16	0.3	0.24
	B-line mean	0.14	0.27	0.21
	R line mean	0.16	0.28	0.23
Ho	Overall range	0-0.21	0-0.41	0-0.3
	Mean	0.02	0.04	0.03
	B-line mean	0.02	0.05	0.04
	R-line mean	0.02	0.03	0.02
He	Overall range	0-0.21	0-0.38	0-0.31
	Mean	0.19	0.38	0.29
	B-line mean	0.16	0.34	0.25
	R-line mean	0.19	0.35	0.27

PIC-Polymorphic information content, Ho- Observed heterozygosity, He- Expected heterozygosity.

10222, ICMR 07666 and ICMR 15333 grouped with B- line clusters and B lines, ICMB 101839, ICMB 101912, ICMB 1502 and ICMA1 19888 grouped with R-line clusters (Figure 3).

The average mean Nei's genetic diversity (GD) among all 373 genotypes was 0.29. Comparing the B- and R-lines, the mean genetic diversity (0.28) of the R lines was higher than the B lines (0.26). Here, 57% of SNPs covered more than the mean GD value among all accessions.

The pair-wise genetic dissimilarity analysis showed that the dissimilarity coefficient for the entire population ranged from 0.003 to 0.68. The minimum and maximum genetic dissimilarly between B- and R-line groups were 0.006 and 0.68, respectively. Within B- and R-line groups, the R-line pairs showed higher genetic dissimilarity (maximum 0.66) over the B-line pairs (maximum 0.58). The total dissimilarity measured in the population was classified into low (0.00 to 0.25), medium (0.25 to 0.50), and high (>0.50) to understand the frequency of pairs present in the respective dissimilarity group. It was observed that 817, 36, and 228 pairs from the B × R, B × B, and R × R groups fell in the high category. On the other hand, 158, 2431, and 1766 pairs were identified as having low genetic dissimilarity under the B × R, B × B, and R × R groups, respectively. It explained that

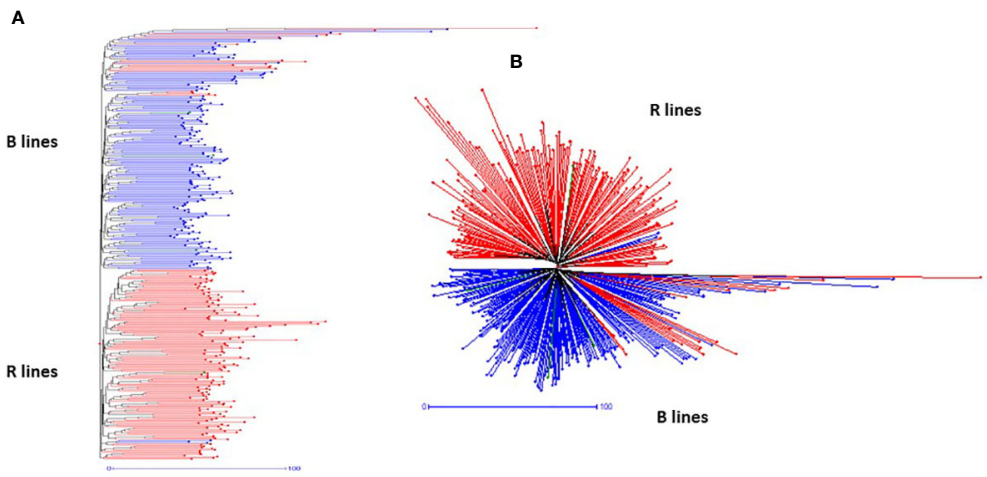


FIGURE 3
Hierarchical **(A)** and radial **(B)** topologies showing the clustering pattern of 373 pearl millet lines based on 4K SNP data.

the diversity among the R × R lines was much higher than the B × B lines.

In the B × R group, genotypes namely, ICMR 100258, ICMR 16333, ICMR 102277, ICMA1 101813, ICMA1 101805, and ICMA1 1803 showed significant genetic diversity over others as they frequently occurred under the high-dissimilarity level. ICMB 92888, ICMB 92111, ICMR 13666, ICMA4 03111, ICMA1 18888, and ICMA1 11999 displayed a high level of genetic relatedness, as the dissimilarity coefficient was the lowest among other pairs. Based on the frequency of occurrence, ICMB 101791 and ICMA1 1803 (within B-lines) and ICMR 100258 and ICMR 16333 (within R-lines) were the genotypes showing a high level of genetic diversity with the other genotypes in their respective groups.

Principal components were generated for the 373 genotypes using the function SnpgdsPCA available in the SNPRelate R package. The percentage of variation calculated for the first 15 principal components was 34%, and the three-dimensional plotting of genotypes was done for the first three components (Figure 4) to determine the grouping pattern of the genotypes. The PCA grouped all the genotypes into two major groups, namely, B and R lines, according to their pedigree which agreed with the genetic diversity model results.

We further investigated the presence of population structure in the 373 genotypes using Structure v2.3.4. The result showed that using mean LnP(K) and delta K values, population structure analysis reveals that the best-assumed group for the current population is 2 (Figure 5). The first sub-population (red color) had 193 genotypes, of which 185 were B-lines and eight were R-lines. A total of 184 genotypes, including 179 R-lines and five B-lines, made up the second sub-population cluster (green color). The genotypes of the B-line cluster (182) and the R-line cluster (108) lie in the > (70-90%) allele frequency range. About 23% of genotype accessions from both population clusters showed some admixtures. The structure model separated the whole population according to the pedigree and matched with the results of GD and PCA models.

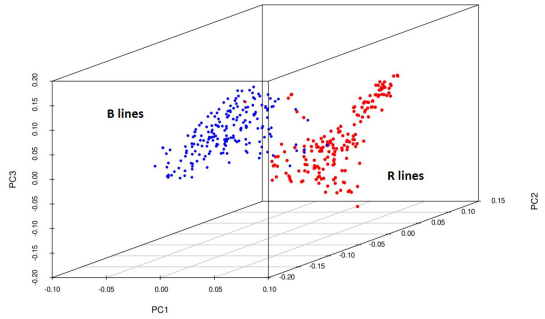


FIGURE 4
3D plot of the principal components from 4K SNP data set explaining a clear-cut separation B- and R-line groups of pearl millet.

Genetic variation among and within group of accessions

AMOVA was performed to estimate the genetic differentiation of populations within and between B- and R-line groups. The results showed that a significant difference was available between the B- and R-line groups. The variation among and within the B- and R-line groups accounted for 13% and 87% of the total variation, respectively. The estimated pairwise PhiPT (Analog of fixation index FST) value between the B- and R-line groups was 0.13. (Table 2).

Linkage disequilibrium

The *r2* was used to estimate LD between all SNPs of the 4K set on each chromosome through TASSEL 5 (Bradbury et al., 2007). Among seven chromosomes, Chr3 showed the highest LD, followed by Chr4, Chr5, and Chr6, while Chr7 showed the lowest LD (Figure 6).

Based on the r^2 value, the LD level was classified as high (0.90-1.00), moderate (0.50-0.90), low (0.10-0.50), and very low (0.00-0.01). Around 278 pairwise SNP were classified as high LD (r^2 >0.90) across the genome. Around 113 SNP pairs showed high LD (0.90-1) at chr3, followed by Chr2, and chr6, which had 47 and 35 high LD SNP pairs, respectively. Clusters of SNP pairs in high LD were primarily found in Chr3, followed by Chr2 and Chr6. A high LD block spanning the length of ~1.5 Mb in the middle of Chr3 (Figure 6) where three SNP pairs covered 1.5Mb distance and 4 SNP pairs covered 291kb. Chr1 had 14 pairs of SNPs between 0.90 - 1 r^2 range, the least among all chromosomes. The length of high LD regions ($r^2 \geq 0.90$) ranged from 3 bp to ~117 Mb. More than 2000 SNP pairs were observed in a moderate LD range (0.50 to 0.90). Of this, Chr3 had the highest pairs (1123) followed by Chr4 (349) and Chr2 (248). The length of the LD blocks was identified in the range of ~27 bp to ~150.93 Mb under the moderate LD category. Around 28K pairs of SNPs have demonstrated low LD (0.10-0.50). Among all, the highest number of pairs was found in Chr3 (15,154 pairs) followed by Chr2 (8,916 pairs) and Chr5 (7037 pairs). The length of the genomic region in the low LD range was extended from 1 Mb to ~300 Mb. More than 167K pairs of SNPs showed a very low LD of <0.10. Chr2 had a high number of LD pairs (~33K), followed by Chr1 (29K) and Chr4 (17K) under the low LD class.

Discussion

Plant breeding dynamics have changed since the 1980s with the development of molecular marker technologies (Hamrick, 1989). Marker-based genetic map facilitates plant genetics and breeding programs that bear the information for desired genes, alleles, or haplotypes. In pearl millet, Liu et al. (1994) created the first genetic map based on 181 RFLP markers. Later, (Qi et al., 2004) identified 353 RFLPs and 65 SSRs for linkage mapping.

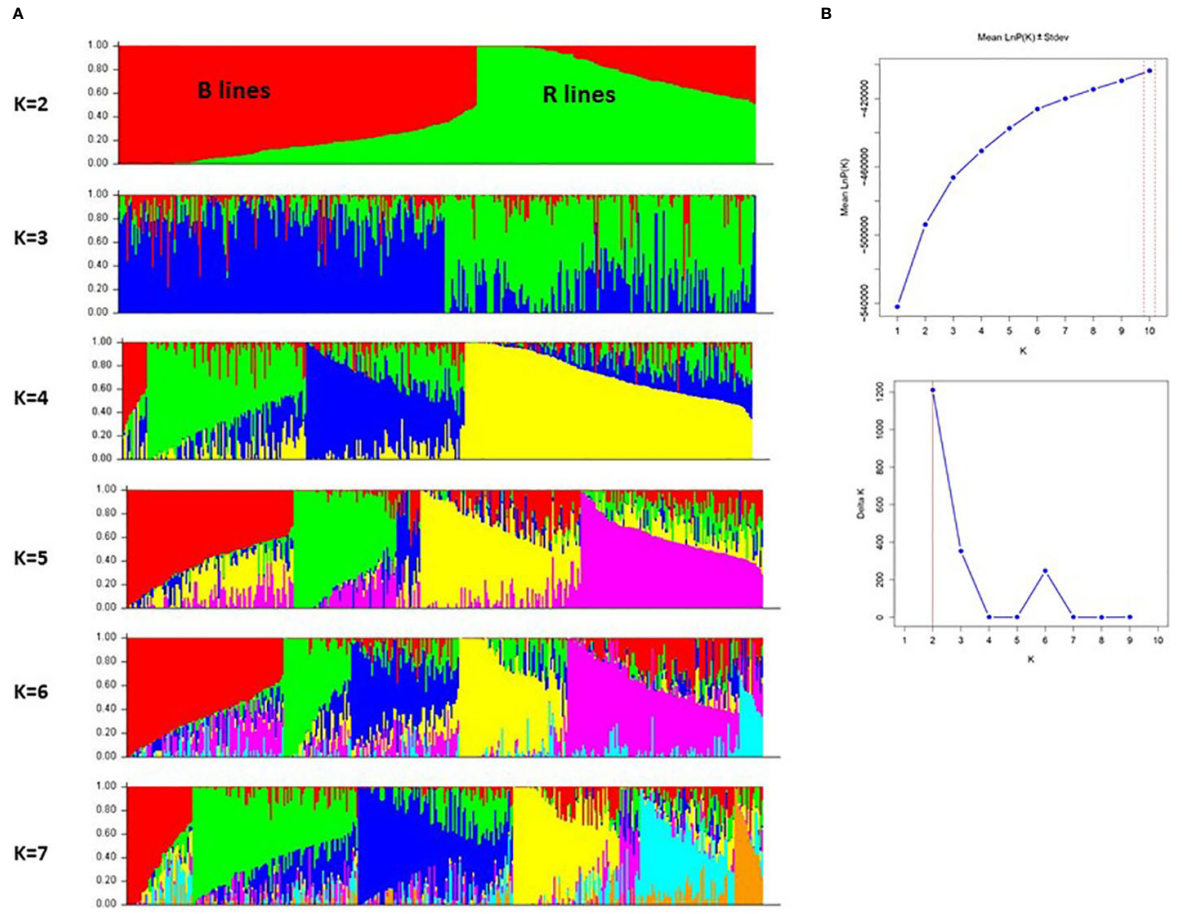


FIGURE 5
(A) A graphical display of the genetic structure of 373 pearl millet inbreds at K value 2 to 7 forming different clusters and exhibiting different levels of admixture. **(B)** Population structure analysis with mean LnP(K) and delta K values, showing the best assumed groups for the given population is 2.

Further, other linkage maps were developed using EST-SSRs and DArT markers (Senthilvel et al., 2008; Supriya et al., 2011). These linkage maps suffer a high degree of marker clustering and lack uniform coverage.

Over the last decade, SNPs have been generated in several crops and have become the most popular genetic marker in trait mapping, molecular breeding, and population genetics experiments. The real explanation for SNPs becoming the

TABLE 2 Analysis of molecular variance of 4K SNPs for the 195 B-lines and 182 R-lines of pearl millet.

Data set	Source	df	SS	MS	Est. Var.	Variance (%)	PhiPT	Significant
4K SNPs	Among groups	1	13226.23	13226.23	67.87	13	0.132	0.001
	Within groups	375	168008.8	448.02	448.02	87		
	Total	376	181235.1		515.89	100		

df- degrees of freedom, SS- Sum of Squares, MS- Mean square, Est.Var- Variance Estimate, PhiPT- Analog of fixation index (Fst).

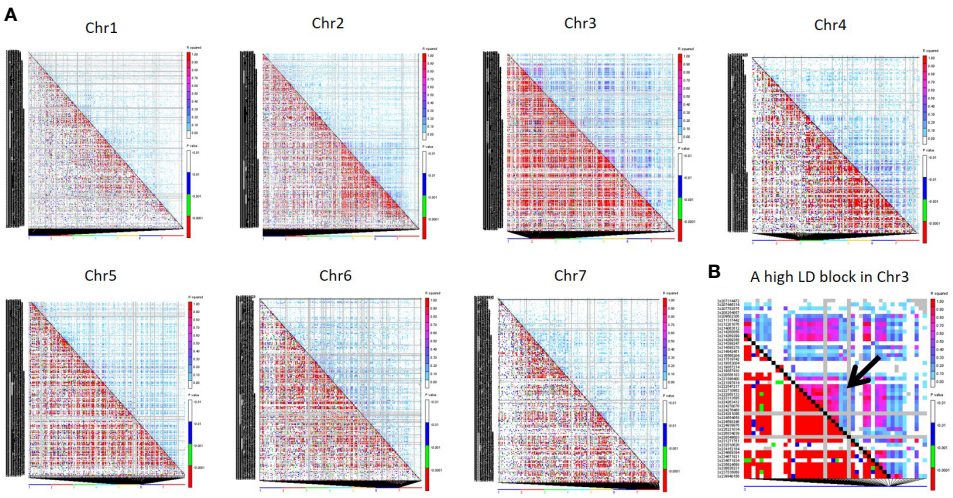


FIGURE 6
(A) Triangle plot of all seven chromosomes representing pairwise LD among all 4K SNPs. Pairwise LD values are plotted on both the X- and Y-axes; above the diagonal displays squared correlation coefficient (r^2) value and below the diagonal displays the corresponding P-values. **(B)** The LD plot of a Chr3 region displays a high LD pattern in a distance of 6.3 Kb to 1.5 Mb for 7 SNP pairs.

marker of choice is their high abundance, even distribution in coding and non-coding regions, and the bi-allelic nature corresponding to the models studied in population genetics and co-dominant mode of inheritance (Khlestkina and Salina, 2006). Additionally, SNP genotyping offers logistical advantages such as a lower rate of genotyping error and greater ease of automating large-scale genotyping (Elshire et al., 2011).

Recent advances in NGS technologies have transformed the pace and precision of plant genomics, providing low-cost genotyping platforms that enable SNPs to be more readily used (Ganal et al., 2012). GBS is an NGS-based SNP detection method to perceive genome-wide SNPs and perform genotyping studies (Elshire et al., 2011). GBS technology provides large data volumes to a range of agronomically essential crops at a low cost per data point, irrespective of previous knowledge of genetic information, genome size, or ploidy information (Scheben et al., 2018).

SNP genotyping with precise sample tracking, collection, and DNA extraction is a potent tool to reshape breeding programs and increase selection gain (Chen et al., 2002). Despite the advantage of having less bias, the high-density SNP genotyping platforms require a sophisticated and complex pipeline with a deep understanding of bioinformatics to analyze the data, which limits the applicability of NGS in many breeding programs (Chen et al., 2014). On the other hand, the mid-density SNP markers are adequate for many breeding experiments at substantially lower cost and complication. Using SNP markers, a 1K RiCA mid-density panel was developed in rice (Arbelaez et al., 2019). Mid-density markers from DArTAG platform were available in maize (3305 SNPs), pigeon pea (2000 SNPs), wheat (3900 SNPs), common bean (1861 SNPs), groundnut (2500 SNPs), cowpea (2602 SNPs), and potato (2147 SNPs) (https://excellenceinbreeding.org). In pearl millet, the *de-novo* genome sequencing, followed by reference-based sequencing of 925 accessions (Varshney et al., 2017), paves the way for developing new SNP tools. Among different SNP densities, a medium density is a worthy addition to the genomic toolbox in pearl millet. Since there is no medium-density genotyping platform available in pearl millet, we developed a viable, cost-effective, and robust 4K SNP panel through "AgriSeq" genotyping technology and demonstrated its functional utility in genetic studies using 373 B and R lines.

The SNP markers overview

The newly developed two 2K sets of SNPs, namely sets I and II, identified independently from the whole-genome sequences PMiGAP panel (345 genotypes) and breeding lines (580 genotypes), respectively, were distributed uniformly across the pearl millet chromosomes (Figure 2A). The gene ontology results showed that 2004, 1646, and 968 SNPs were associated with molecular function, biological process, and cellular component, respectively. Among various molecular functions, ATP binding was associated with more SNPs (318) followed by protein binding (235) and DNA binding (166) classes. More than 30 SNPs were associated with abiotic, oxidative, salt, and osmotic stresses under the biological process category. Nine SNPs identified in the cellular process category were related to heat shock responses. Membrane and nuclear-related functions were the top ones, as 38% of SNPs captured them from the cellular component.

The polymorphism information content (PIC) is one of the important measures to calibrate the informativeness of the marker. The higher value indicated that a marker has more alleles and can discriminate most individuals in a population (Botstein et al., 1980). Markers in our experiment had a PIC value as high as 0.38. The 2K set II had a high PIC (0.3) than 2K set I (0.16), while the combined 4K set had an average PIC of 0.23. The average expected heterozygosity (He) was higher (0.29) than the observed heterozygosity (Ho) (0.03) in the 4K set. The low level of observed heterozygosity was attributed to the fact that the inbreds attained almost homozygosity across loci, with minimal residual heterozygosity in the population. Among B- and R-lines, the observed heterozygosity in B-lines was higher than over R-lines. The selected marker panel genotyped through "AgriSeq" technology proved its utility as it discriminated the homozygotes and heterozygotes.

Genetic diversity, PCA, and population structure

The pattern and degree of genetic diversity among 373 genotypes representing the gene pool of B and R groups were examined using 4K set SNPs. All genotypes were divided into two major groups based on the NJ analysis of the genetic dissimilarity coefficients. The B- and R- lines were further grouped into three subgroups, with some of the B-lines clustered with R-lines and *vice-versa* due to the fact that they share some level of parentage with the respective neighbors. The cross-grouping of B and R lines was also found in previous studies in pearl millet (Nepolean et al., 2012). Pairs with different levels of genetic dissimilarity were identified in both B and R groups. R groups had higher mean genetic diversity over B-lines. The higher gene diversity and the more alleles detected in R-lines were attributed to the broader genetic base of these lines while breeding these pollinator lines. The genetic distance among the lines will provide an opportunity to select precise crosses in heterosis breeding programs (Thirunavukkarasu et al., 2013). A set of highly genetically dissimilar pairs were identified within the B- and R-line pools. New B- and R-lines can be created by exploiting the genetic variability available in the respective pools (Table 3). Additionally, pairs with high genetic dissimilarity between B- and R-line pools were identified. These pairs can be used for generating heterotic combinations by exploiting the GCA and other beneficial agronomic traits (Table 4).

The number of subpopulations was validated by plotting the PCA of the genetic data. PCA captures the continuous axes of genetic variation by correlating and ranking the genotypes (Price et al., 2006). The PCA showed two major groups plotted in the first three axes. There is great diversity within these subgroups, as evidenced by the fact that the first 15 PCA components explained more than 34% of the variation and broader divergence of the lines. Population structure plays an integral part in understanding evolutionary genetics and illustrating the diversity of a population. In the present study, the structure model revealed that the 373 genotypes from diverse sources originated from two genetic populations (K= 2), which were expected as they belong to different B and R breeding groups. Population structure also revealed a smaller amount of admixture between the two populations, which explained that they share common breeding history. While developing the B and R lines, the lines from other groups might have been used to introduce new and valuable traits unavailable in the respective pools.

The grouping of B and R lines in our study can be traced back to the history of breeding in pearl millet. The B- and R-lines are named female lines and pollinator lines, respectively, in hybrid breeding. These two groups represent a putative heterotic genotype pool having favorable alleles for increasing the yield. In order to maintain the heterotic potential between B and R lines, line development programs strictly used the lines within respective groups, and new hybrids were generated using B and R line crosses. Hence, separate genetic pools have been maintained between two different populations. Our newly developed 4K SNP panel captured the genetic properties such as genetic diversity, PCA, and population structure of the B and R lines. The genetic relationship between and among B and R line groups will be helpful in developing new heterotic pools and segregating populations for trait mapping, and conducting association mapping and genomic selection experiments.

Genetic variation among and within group of accessions

The B and R line groups, classified based on the pedigree, were analyzed to characterize the genetic differentiation between and within the subgroups using AMOVA. The relative contribution between populations to the overall genetic variation is described by phi-statistics (PhiPT), a modified form of Wright's F(Fst). The genetic variation between individuals within a population and the population's divergence from the Hardy-Weinberg proportions are measured by Fst (Wright et al., 1978). The AMOVA results of the current experiment showed that the majority of the variation within sub-populations accounted for 87% ($P < 0.001$) of the total variation, and the between-population differences accounted for 13% ($P < 0.001$) of the variation (Table 2). It indicated that a large part of the accessions within the groups showed a high-level genetic variability. Previously, a set of 213 old and 166 newly-generated pearl millet parental lines were genotyped by 28 SSRs, and the subsequent AMOVA analysis of the old and new sets showed that the genetic variation between B and R lines was 16.98% and 9.22%, respectively (Gupta et al., 2015).

This was further supported by the combined AMOVA of both sets, which showed a significant difference between the B- and R-line groups. A range of 0 to 0.05, 0.05 to 0.15, 0.15 to 0.25, and >0.25 indicate little, moderate, large, and great genetic differences, respectively (Wright et al., 1978). While comparing the genetic

TABLE 3 Pairs of highly genetically dissimilar (>0.45) lines captured by the 4K SNP panel for use in developing new lines in respective B- and R-line pools.

S.No	B-lines			R-lines		
	Line 1	Line 2	Genetic dissimilarity value	Line 1	Line 2	Genetic dissimilarity value
1	ICMB 101791	ICMA1 101805	0.578	ICMR 100258	ICMR 07777	0.655
2	ICMA1 1803	ICMB 101564	0.545	ICMB 92888	ICMR 100258	0.592
3	ICMA1 101877	ICMA1 101813	0.504	ICMB 92111	ICMR 100258	0.580
4	ICMB 101578	ICMA1 1803	0.499	ICMR 100258	ICMR 19222	0.579
5	ICMA1 09222	ICMA1 101813	0.495	ICMR 102545	ICMR 16333	0.565
6	ICMA1 101813	ICMA1 101799	0.493	ICMR 100258	ICMR 1203	0.565
7	ICMB 101830	ICMA1 101813	0.492	ICMR 101859	ICMR 100258	0.563
8	ICMA1 101327	ICMB 101564	0.486	ICMR 16333	ICMR 14111	0.547
9	ICMB 101888	ICMB 101791	0.485	ICMB 92888	ICMR 16333	0.540
10	ICMB 101793	ICMB 101791	0.481	ICMR 102277	ICMR 14111	0.527
11	ICMB 101831	ICMB 101600	0.480	ICMR 11999	ICMR 07777	0.521
12	ICMB 101578	ICMA4 02111	0.476	ICMR 102283	ICMR 16333	0.520
13	ICMA5 02444	ICMA4 01444	0.476	ICMR 07777	ICMR 102539	0.520
14	ICMA1 18888	ICMB 101564	0.475	ICMR 18888	ICMR 16333	0.520
15	ICMB 101578	ICMA1 92777	0.475	ICMR 17111	ICMR 16333	0.518
16	ICMB 101879	ICMB 101564	0.473	ICMR 14111	ICMR 10888	0.514
17	ICMA1 101813	ICMA1 101811	0.473	ICMR 19888	ICMR 16333	0.514
18	ICMB 101878	ICMA1 101813	0.472	ICMR 16333	ICMR 101307	0.512
19	ICMA1 101805	ICMA1 101799	0.471	ICMB9 2111	ICMR 102277	0.511
20	ICMB 101832	ICMA1 101805	0.468	ICMR 13999	ICMR 07777	0.503
21	ICMB 101889	ICMB 101791	0.468	ICMR 19222	ICMR 16333	0.503
22	ICMA1 92777	ICMA1 100718	0.467	ICMR 09222	ICMR 07777	0.502
23	ICMB 101791	ICMB 101789	0.466	ICMR 102012	ICMR 16333	0.502
24	ICMB 101885	ICMA5 02444	0.464	ICMR 102547	ICMR 07777	0.502
25	ICMA1 19888	ICMB 101564	0.463	ICMR 16333	ICMR 102504	0.502

differentiation, rice showed PhiPT value of 0.130 between the *indica* and *japonica* groups (Luong et al., 2021), and the Ethiopian sorghum group showed 0.252 between B- and R-lines (Mindaye et al., 2015). Finger millet revealed a moderate genetic differentiation (Fst = 0.352) among seven population sub-groups (Brhane et al., 2022). Pearl millet demonstrated genetic differentiation at a PhiPT value of 0.130 for the chosen genotypes, similar to those studies. It also implied that the current marker set used in this experiment could extract the molecular variance at a population level and can be used for further applications such as designing crosses based on genetic diversity, developing mapping populations for trait mapping, and conducting genomic selection experiments.

Linkage disequilibrium

The potential response to both natural and artificial selection is constrained by the non-random association of alleles at two or more loci, which also offers information about past events. LD reflects the history of natural selection, gene conversion, mutation, and other forces influencing gene frequency and evolution. LD provides insight into previous evolutionary events and explains the co-evolution of linked sets of genes. LD-based on Pearson correlations (r^2) is a squared value of the correlation between pairs of markers across the genome. The details of the LD pattern are used in mapping genes associated with complex quantitative traits. In association studies, it has frequently been discovered that

TABLE 4 Pairs of highly genetically dissimilar (>0.5) lines captured by the 4K SNP panel for use in developing new hybrid combinations.

S No	Potential crosses		Genetic dissimilarity Value
	B-Line	R-Line	
1	ICMA1 101813	ICMR 100258	0.687
2	ICMB 101791	ICMR 16333	0.606
3	ICMB 101791	ICMR 07777	0.595
4	ICMB 101791	ICMR 16777	0.553
5	ICMB 101791	ICMR 10888	0.568
6	ICMB 101791	ICMR 102497	0.563
7	ICMA5 02444	ICMR 102545	0.560
8	ICMB 101791	ICMR 14111	0.546
9	ICMB 101791	ICMR 16444	0.543
10	ICMB 101791	ICMR 102277	0.543
11	ICMB 101791	ICMR 10111	0.538
12	ICMB 101791	ICMR 100168	0.530
13	ICMA1 101813	ICMR 17333	0.530
14	ICMB 101791	ICMR 102151	0.530
15	ICMA1 101813	ICMR 102499	0.529
16	ICMA1 101813	ICMR 07111	0.529
17	ICMB 101791	ICMR 08888	0.526
18	ICMA1 101813	ICMR 07888	0.525
19	ICMA1 101813	ICMR 18111	0.524
20	ICMB 101791	ICMR 15333	0.524
21	ICMA1 101813	ICMR 06999	0.524
22	ICMB 101791	ICMR 12111	0.524
23	ICMA1 101813	ICMR 102548	0.523
24	ICMB 101791	ICMR 07999	0.522
25	ICMA1 101813	ICMR 15111	0.521

markers directly related to the mutation exhibit less LD than those more distantly related. The test of LD is crucial because it will help to quickly and efficiently choose the SNP markers that can be used for trait mapping and selection studies.

In our result, uniformly distributed markers showed the regions of high and low LD on various chromosomes of pearl millet (Figure 6). Comparing the high LD pairs on all chromosomes, Chr3 captured 40% high LD pairs, and the length of the high-LD pairs on Chr3 ranged from 1Mb to 117Mb. Chr3 is the longest one (346 Mb) among all chromosomes, so that it would have captured more LD events. The higher LD in Chr3 is attributed to the low level of recombination events and fixation of alleles. Among high LD (r^2 <0.90) SNP pairs, 63% were derived from PMiGAP, and 36% SNPs from breeding lines since the PMiGAP represented the accessions with greater genetic diversity over the cultivated breeding lines and was clearly captured by the newly developed 4K SNP panel. The knowledge of LD regions from the B and R lines of the pearl millet genomes characterized by the newly developed mid-density set provides the opportunity to exploit them in the genetic characterization of diverse germplasm, trait mapping, and genomic selection experiments.

Conclusions

We have successfully identified and validated a mid-density marker set for routine genotyping of pearl millet lines and its usefulness for various genomic studies. By

mining 925 pearl millet genomes comprising genetically diverse wild and breeding lines, a set of 4112 SNPs were identified. A panel of 373 B and R lines was genotyped by these SNPs using the Agri-Seq platform. The results showed that the newly developed SNPs were uniformly distributed across the genome and had significant PIC and gene diversity. The SNP panel was used to group the genotypes through diversity, PCA, and structure models. All three statistics showed consistent results where they separated the accessions into two major groups, B and R lines. The LD analysis showed the regions of high and low LD. The AMOVA revealed a significant distinction between the B- and R-line groups and the extent of genetic divergence within and across and R lines groups. This research demonstrated that pearl millet has a high degree of genetic diversity and variable levels of LD across the genome, which are highly beneficial in developing heterotic groups for hybrid breeding. The experiment revealed that our mid-density 4K SNP panel genotyped by AgriSeq technology had a high level of information, making them suitable for several uses, including trait mapping, marker-assisted backcrossing, and genomic selection for pearl millet improvement.

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.

Author contributions

NT conceptualized the experiment. JS, SS, AR, AC, JI performed the data analysis. SKG, RM, MG, SR contributed and maintained the genotypes. KG, SH, AV performed the genotyping experiments. All authors contributed to the final manuscript. All authors read and approved the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The experiment was funded by the Bill and Melinda Gates Foundation project (INV-008187) and the ICAR-Indian Institute of Millets Research project (CI/2018-23/120).

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.2022.1068883/full#supplementary-material

References

Ambawat, S., Senthilvel, S., Hash, C. T., Nepolean, T., Rajaram, V., Eshwar, K., et al. (2016). QTL mapping of pearl millet rust resistance using an integrated DArT-and SSR-based linkage map. *Euphytica* 209 (2), 461–476. doi: 10.1007/s10681-016-1671-9

Arbelaez, J. D., Dwiyanti, M. S., Tandayu, E., Llantada, K., Jarana, A., Ignacio, J. C., et al. (2019). 1k-RiCA (1K-rice custom amplicon) a novel genotyping amplicon-based SNP assay for genetics and breeding applications in rice. *Rice* 12 (1), 1–15. doi: 10.1186/s12284-019-0311-0

Excellence in breeding platform. Available at: https://excellenceinbreeding.org (Accessed July,2022).

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 (3), 314.

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

Brhane, H., Haileselassie, T., Tesfaye, K., Ortiz, R., Hammenhag, C., Abreha, K. B., et al. (2022). *Novel gbs-based snp markers for finger millet and their use in genetic diversity analyses* Vol. 13 (Frontiers in Genetics). doi: 10.3389/fgene.2022.848627

Chen, M., Presting, G., Barbazuk, W. B., Goicoechea, J. L., Blackmon, B., Fang, G., et al. (2002). An integrated physical and genetic map of the rice genome. *Plant Cell* 14 (3), 537–545. doi: 10.1105/tpc.010485

Chen, H., Xie, W., He, H., Yu, H., Chen, W., Li, J., et al. (2014). *A high-density SNP genotyping array for rice biology and molecular breeding* (Mol Plant 7:541–553). doi: 10.1093/mp/sst135

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

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6, e19379. doi: 10.1371/journal.pone.0019379

Ganal, M. W., Plieske, J., Hohmeyer, A., Polley, A., and Röder, M. S. (2019). "High-throughput genotyping for cereal research and breeding," in *Applications of genetic and genomic research in cereals* (Woodhead Publishing), pp3–p17. doi: 10.1016/B978-0-08-102163-7.00001-6

Ganal, M. W., Polley, A., Graner, E. M., et al. (2012). Large SNP arrays for genotyping in crop plants. *J Biosci* 37, 821–828. doi: 10.1007/s12038-012-9225-3

Granato, I. S., Galli, G., de Oliveira Couto, E. G., Mendonça, L. F., and Fritsche-Neto, R. (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38 (8), 1–7. doi: 10.1007/s11032-018-0844-8

Gupta, S. K., Nepolean, T., Sankar, S. M., Rathore, A., Das, R. R., Rai, K. N., et al. (2015). Patterns of molecular diversity in current and previously developed hybrid parents of pearl millet [Pennisetum glaucum (L.) r. br.]. *Am. J. Plant Sci.* 6 (11), 1697–1712. doi: 10.4236/ajps.2015.611169

Hamrick, J. L., and Godt, M. W.. (1996). Effects of life history traits on genetic diversity in plant species. Philosophical Transactions of the Royal Society of London. . *Series B: Biological Sciences* 351 (1345), 1291–1298.

Kanatti, A., Rai, K. N., Radhika, K., Govindaraj, M., Sahrawat, K. L., Srinivasu, K., et al. (2014). Relationship of grain iron and zinc content with grain yield in pearl millet hybrids. *Crop Improv.* 41, 91–96. Availabe at: http://oar.icrisat.org/id/eprint/8894.

Keneni, G., Bekele, E., Imtiaz, M., Dagne, K., Getu, E., and Assefa, F. (2012). Genetic diversity and population structure of Ethiopian chickpea (Cicer arietinum l.) germplasm accessions from different geographical origins as revealed by microsatellite markers. *Plant Mol. Biol. Rep.* 30 (3), 654–665. doi: 10.1007/s11105-011-0374-6

Khlestkina, E. K., and Salina, E. A. (2006). SNP markers: Methods of analysis, ways of development, and comparison on an example of common wheat. *Russ J. Genet.* 42, 585–594. doi: 10.1134/S1022795406060019

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). Diversity arrays technology: a generic genome profiling technology on open platforms. *Methods Mol. Biol.* 888, 67–89. doi: 10.1007/978-1-61779-870-2_5

Koelewijn, H. P. (2018). "Advancing vegetable breeding with applied biosystems™ AgriSeq™ targeted genotyping by sequencing (GBS)," In *Plant and Animal Genome XXVI Conference*, (January 13-17, 2018). PAG.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15 (5), 1179–1191. doi: 10.1111/1755-0998.12387

Kumar, S., Hash, C., Nepolean, T., Mahendrakar, M., Satyavathi, C., Singh, G., et al. (2018). Mapping grain iron and zinc content quantitative trait loci in an iniadi-derived immortal population of pearl millet. *Genes* 9, 248. doi: 10.3390/genes9050248

Kumar, S., Hash, C. T., Thirunavukkarasu, N., Singh, G., Rajaram, V., Rathore, A., et al. (2016). Mapping quantitave trait loci controlling high iron and zinc in self and open pollinated grains of pearl millet [Pennisetum glaucum (L) r. *Br. Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01636

Liu, C. J., Witcombe, J. R., Pittaway, T. S., Nash, M., Hash, C. T., Busso, C. S., et al. (1994). An RFLP-based genetic map of pearl millet (Pennisetum glaucum). *Theoret. Appl. Genet.* 89 (4), 481–487. doi: 10.1007/BF00225384

Luong, N. H., Linh, L. H., Shim, K. C., Adeva, C., Lee, H. S., and Ahn, S. N. (2021). Genetic structure and geographical differentiation of traditional rice (Oryza sativa l.) from northern Vietnam. *Plants* 10 (10), 2094. doi: 10.3390/plants10102094

Mindaye, T. T., Mace, E. S., Godwin, I. D., and Jordan, D. R. (2015). Genetic differentiation analysis for the identification of complementary parental pools for sorghum hybrid breeding in Ethiopia. *Theor. Appl. Genet.* 128 (9), 1765–1775. doi: 10.1007/s00122-015-2545-6

Nambiar, V. S., Dhaduk, J. J., Sareen, N., Shahu, T., and Desai, R. (2011). Potential functional implications of pearl millet (Pennisetum glaucum) in health and disease. *J. Appl. Pharm. Sci.* 01, 62–67. doi: 10.1105/tpc.109.068437

Nepolean, T., Firoz, H., Kanika, A., Rinku, S., Kaliyugam, S., Swati, M., et al. (2014). Functional mechanisms of drought tolerance in subtropical maize (Zea mays l.) identified using genome-wide association mapping. *BMC Genomics* 15 (1182), 1–12. doi: 10.1186/1471-2164-15-1182

Nepolean, T., Gupta, S. K., Dwivedi, S. L., Bhattacharjee, R., Rai, K. N., and Hash, C. T. (2012). Genetic diversity in maintainer and restorer lines of pearl millet. *Crop Sci.* 52, 2555–2563. doi: 10.2135/cropsci2011.11.0597

Peakall, R. O. D., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in excel. population genetic software for teaching and research. *Mol. Ecol. Notes* 6 (1), 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Perrier, X., and Jacquemoud-Collet, J. P. (2006) *DARwinSoftware*. Available at: http://darwin.cirad.fr/darwin (Accessed July, 2022).

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, 945–959. doi: 10.1093/genetics/155.2.945

Qi, X., Pittaway, T. S., Lindup, S., Liu, H., Waterman, E., Padi, F. K., et al. (2004). An integrated genetic map and a new set of simple sequence repeat markers for pearl millet, pennisetum glaucum. *Theor. Appl. Genet.* 109, 1485–1493. doi: 10.3835/plantgenome2012.06.0006

Rajaram, V., Nepolean, T., Senthilvel, S., Varshney, R. K., Vadez, V., Srivastava, R. K., et al. (2013). Pearl millet [Pennisetum glaucum (L.) r. br.] consensus linkage map constructed using four RIL mapping populations and newly developed EST-SSRs. *BMC Genomics* 14, 159. doi: 10.1007/s00122-004-1765-y

Ren, R., Ray, R., Li, P., Xu, J., Zhang, M., Liu, G., et al. (2015). Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. *Mol. Genet. Genomics* 290, 1457–1470. doi: 10.1007/s00438-015-0997-7

Scheben, A., Batley, J., and Edwards, D. (2018). *Revolution in genotyping platforms for crop improvement*. In R. Varshney, M. Pandey and A. Chitikineni (eds) *Plant Genetics and Molecular Biology. Advances in Biochemical Engineering/Biotechnology* (Berlin, Heidelberg: Springer) 164. doi: 10.1007/10_2017_47

Sehgal, D., Rajaram, V., Armstead, I. P., Vadez, V., Yadav, Y. P., Hash, C. T., et al. (2012). Integration of gene-based markers in a pearl millet genetic map for identification of candidate genes underlying drought tolerance quantitative trait loci. *BMC Plant Biol.* 12 (1), 9. doi: 10.1007/s00122-013-2197-3

Senthilvel, S., Jayashree, B., Mahalakshmi, V., Kumar, P. S., Nakka, S., Nepolean, T., et al. (2008). Development and mapping of simple sequence repeat markers for pearl millet from data mining of expressed sequence tags. *BMC Plant Biol.* 8, 119. doi: 10.1371/journal.pone.0122165

Senthilvel, S., Nepolean, T., Supriya, A., Rajaram, V., Kumar, S., Hash, C. T., et al. (2010). "Development of a molecular linkage map of pearl millet integrating DArT and SSR markers," in *Plant and Animal Genome Conference*, San Diego, CA. 9–13. doi: 10.1186/1471-2229-8-119

Supriya, A., Senthilvel, S., Nepolean, T., Eshwar, K., Rajaram, V., Shaw, R., et al. (2011). Development of a molecular linkage map of pearl millet integrating DArT and SSR markers. *Theor. Appl. Genet.* 123, 239–250. doi: 10.1371/journal.pgen.1004982

Thirunavukkarasu, N., Hossain, F., Shiriga, K., Mittal, S., Arora, K., Rathore, A., et al. (2013). Unraveling the genetic architecture of subtropical maize (Zea maysL.) lines to assess their utility in breeding programs. *BMC Genomics* 14 (1), 1–13. doi: 10.1186/1471-2164-14-877

Thirunavukkarasu, N., Sharma, R., Singh, N., Shiriga, K., Mohan, S., Mittal, S., et al. (2017). Genomewide expression and functional interactions of genes under drought stress in maize. *Int. J. Genomics* 2017, 1–13. doi: 10.1155/2017/2568706

Thomson, M. J., Zhao, K., Wright, M., McNally, K. L., Rey, J., Tung, C. W., et al. (2012). High-throughput single nucleotide polymorphism genotyping for breeding applications in rice using the BeadXpress platform. *Mol. Breed.* 29 (4), 875–886. doi: 10.1007/s11032-011-9663-x

Varshney, R. K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., et al. (2017). Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nat. Biotechnol.* 35 (10), 969–976. doi: 10.1038/nbt.3943

Wright, S. (1978). *Evolution and the genetics of populations, volume 4: variability within and among natural populations (Vol. 4)* (University of Chicago press).

Wright, S. (1978). *) Evolution and the Genetics of Populations. Vol. 4. Variability within and among Natural Populations* (Chicago: University of Chicago press).

Wu, X., Li, Y., Shi, Y., Song, Y., Wang, T., Huang, Y., et al. (2014). Fine genetic characterization of elite maize germplasm using high-throughput SNP genotyping. *Theor. Appl. Genet.* 127, 621–631. doi: 10.1007/s00122-013-2246-y

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28 (24), 3326–3328. doi: 10.1093/bioinformatics/bts606

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 16 January 2023
DOI 10.3389/fpls.2022.1079571

OPEN ACCESS

EDITED BY
Patricio Hinrichsen,
Agricultural Research Institute, Chile

REVIEWED BY
Kevin Kit Siong Ng,
Forest Research Institute Malaysia (FRIM), Malaysia
Heino Konrad,
Austrian Research Centre for Forests (BFW), Austria

*CORRESPONDENCE
Hanguo Zhang
hanguozhang1@sina.com

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

RECEIVED 25 October 2022
ACCEPTED 29 December 2022
PUBLISHED 16 January 2023

CITATION
Yan P, Xie Z, Feng K, Qiu X, Zhang L and Zhang H (2023) Genetic diversity analysis and fingerprint construction of Korean pine (*Pinus koraiensis*) clonal seed orchard. *Front. Plant Sci.* 13:1079571. doi: 10.3389/fpls.2022.1079571

Genetic diversity analysis and fingerprint construction of Korean pine (*Pinus koraiensis*) clonal seed orchard

Pingyu Yan[1], Zixiong Xie[1], Kele Feng[1], Xinyu Qiu[2], Lei Zhang[1] and Hanguo Zhang[1]*

[1]State Key Laboratory of Tree Genetics and Breeding, Northeast Forestry University, Harbin, China, [2]Heilongjiang Academy of Forestry, Harbin, China

Korean pine is a native tree species in Northeast China. In order to meet the needs of germplasm resource evaluation and molecular marker-assisted breeding of Korean pine, we collected Korean pine clones from 7 populations in Northeast China, analyzed the genetic diversity and genetic structure by SSR molecular marker technology and clustered them to revealed the inter- and intrapopulation differentiation characteristics of each clone. The fingerprint profiles of 161 Korean pine clones were also constructed. 77 alleles were detected for 11 markers, and 18 genotypes were identified on average for each marker. The PIC of the different markers ranged from 0.155-0.855, and the combination of PI and PIsibs for the 11 markers was 3.1×10^{-8} and 1.14×10^{-3}, respectively. MANOVA showed that genetic variation existed mainly within populations, accounting for 98% of the total variation. The level of genetic differentiation among populations was low, with an average *Nm* between populations of 11.036. Genetic diversity is lower in the Lushuihe population and higher in the Tieli population. The 161 Korean pine clones were divided into 4 or 7 populations, and the 7 populations were not clearly distinguished from each other, with only the Lushuihe population showing partial differentiation. There is no significant correlation between the genetic distance of Korean pine populations and the geographical distance of their superior tree sources. This result can provide recommendations for future Korean pine breeding programs. The combination of 11 markers could completely distinguish 161 clones and establish the fingerprint. Genetic diversity of Korean pine clones from the 7 populations was abundant, and the genetic distances of individuals and populations were evenly dispersed. The fingerprint map can be used for the identification of Korean pine clones.

KEYWORDS

Korean pine, SSR, genetic diversity, clone, fingerprint

1 Introduction

Genetic diversity is the basis of evolution (Hughes et al., 2008) and provides the raw material for evolution of natural selection (Nevo, 1988; Zhang et al., 2022). Intraspecific genetic variation is the basis and most basic level of biodiversity (Pauls et al., 2013), and it is important for the evolution and conservation of species (Ellegren and Galtier, 2016). The level of genetic diversity within a population affects the productivity, growth and stability of that population (Hughes et al., 2008). Genetic diversity may not necessarily enable a population to persist, but reduced genetic diversity in a population may have long-term effects on its future evolution, as well as on its adaptive capacity in times of environmental change (Jump et al., 2009; Markert et al., 2010). The assessment of genetic diversity within and among populations is important for decision-making in genetic conservation programs, because studying the relationship between genetic diversity and fitness can predict the importance of genetic diversity for a given population (Eding et al., 2002). The genetic basis of a breeding population determines the genetic quality and long-term potential of breeding programs and products (Ivetić et al., 2016). The size of parental populations determines the level of genetic diversity in new stands (Flint-Garcia, 2013), so it is in our best interest to maintain diversity and promote systematic redundancy and resilience (Ledig, 1992). To avoid population genetic bottlenecks and maintain maximum effective population size, appropriate sampling strategies can maximize increase genetic diversity in the population of seed production (Ivetic et al., 2016). Regular monitoring of trends in genetic diversity utilization in breeding programs can provide breeders with options for developing new varieties and hybrids (Govindaraj et al., 2015; Jin et al., 2016).

Korean pine (*Pinus koraiensis*), a genus of pine in the family Pinaceae, National Key Conserved Wild Plants of Grade II in China, is a native species in northeast China (Lim, 2012). Traditionally, Korean pine is a good tree species capable of providing wood, pulp and oil. In addition, the seed of Korean pine is the most popular pine nut due to its nutritional value (Yoon et al., 1989; Wolff et al., 2000), high amounts of crude protein, crude fat, polysaccharides and crude fiber as well as vitamins, minerals and trace elements (Ca, P, Mn, Co, Cu and Zn) (Nergiz and Donmez, 2004; Zadernowski et al., 2009). The market demand for superior Korean pine seeds has promoted the development of Korean pine clones seed orchard, which were established in China as early as the early 1960s, and the technical system for the creation from fringe picking to seedlings management was proposed in the 1970s. Subsequently, Korean pine clones seed orchards were established in many places in northeast China to improve the genetic quality of Korean pine seeds that can be used for afforestation. At the same time, research on productivity techniques, flowering and fruiting patterns in Korean pine seed orchards is also being conducted (An et al., 1992). These excellent Korean pine resources have become important conventional breeding materials and are used in traditional breeding studies, including analysis of fruiting traits, selection of superior clones, analysis of seed traits, nutrient composition, variation studies of seed traits, genetic diversity analysis and studies on phenotypic diversity of needles and cones in Korean pine seed orchard (Zhang et al., 2015b; Tong et al., 2019; Weihuai et al., 2019; Pingyu et al., 2020; Qianping, 2020; Longhai et al., 2021). In addition, studies on the reaction conditions of ISSR, SSR and SRAP in Korean pine, laying the foundation for genetic differentiation of Korean pine populations based on molecular markers (Feng et al., 2004; Feng et al., 2010; Zhao et al., 2010).

Follow-up surveys conducted to confirm clones have generally shown that mislabeling of seed orchard divisions is relatively common (Wheeler and Jech, 1992). Plant varieties are often identified by morphology, traditionally. However, it is difficult to identify different clones morphologically, because there is little morphological variation among clones, and some morphological appearances are susceptible to environmental factors. The limitations of genetic markers for phenotype have led to the development of more effective-directly DNA-based markers called molecular markers, which is specific DNA fragments representing genome-level differences (Agarwal et al., 2008). Microsatellite is ideal for identifying individuals and studying genetic diversity, due to their ubiquity, reproducibility, a high level of polymorphism, co-dominant and high levels of transferability (Guan et al., 2019; Lv et al., 2020; Nn et al., 2020; Carletti et al., 2021). Therefore, SSR has been used for genetic diversity studies, genetic linkage, and fingerprinting of many important economic tree species, such as Date palm (*Ziziphus jujuba* Mill.), Poplar (*Populus* L.), and Pear (*Pyrus* spp) (Liang et al., 2005; Gao et al., 2012; Ma et al., 2012), as well as pines such as Masson pine (*Pinus massoniana*) (Afeng, 2005) and Camphor pine (*Pinus sylvestris* var. *mongolica*) (Huili et al., 2022).

In this study, we collected 161 clones from 7 Korean pine seed orchards in northeastern China. 11 SSR genotyping data of 161 clones of Korean pine were obtained by capillary electrophoresis. The fingerprint map of Korean pine clones was established, which provides a strong guarantee of technology for resource sharing and the distribution application of superior clones, and has important value in property protection and promotion of superior seed. In addition, the genetic diversity and genetic structure of Korean pine clones seed orchard are evaluated and systematically described, which can help improve the utilization efficiency of Korean pine resources, guide the development of further breeding strategies, and provide a basis for the scientific utilization of Korean pine germplasm resources.

2 Materials and methods

2.1 Plant materials and DNA extraction

In this study, a total of 161 clones were collected from 7 Korean pine seed orchards in Heilongjiang and Jilin Province, whose superior tree (refers to individuals with excellent growth, timber and resistance adaptations in natural or planted forests with similar environmental conditions, such as the same stand conditions, the same forest age and the same forestry measures) originated from 6 sites in Changbai Mountains and Xiaoxinganling, the main distribution areas of Korean pine (Table 1). Total of 805 samples collected, with 5 ramets has collected from each clone. Annual conifers of Korean pine were collected and snap-frozen in liquid nitrogen for DNA extraction.

Total DNA of Korean pine samples was extracted using the DP-320 Plant Genome Extraction Kit (Tiangen, Beijing, China). The integrity of genomic DNA was examined using a 1% agarose gel, and DNA concentration and quality were examined using Micro-

TABLE 1 Summary of material source information.

Population	Source of Superior Tree	Location (°)	Elevation (m)	Number of clones	Clones
Bohai	Xiaobeihu	N 44.21; E 128.56	743	18	BH1, BH6, BH8, BH16, BH26, BH38, BH45, BH51, BH61, BH63, BH66, BH67, BH69, BH70, BH71, BH73, BH92, BH93
Hegang	Wuying	N 48.24; E 129.25	547	26	HG3, HG4, HG7, HG8, HG9, HG10, HG11, HG12, HG14, HG15, HG17, HG21, HG24, HG25, HG26, HG27, HG28, HG29, HG30, HG31, HG39, HG40, HG44, HG46, HG47, HG51
Lushuihe	Lushuihe	N 42.47; E 127.78	775	21	LSH21, LSH22, LSH25, LSH38, LSH96, LSH99, LSH105, LSH106, LSH117, LSH127, LSH132, LSH139, LSH161, LSH162, LSH165, LSH169, LSH179, LSH193, LSH194, LSH331, LSH428
Weihe	Hebei	N 48.08; E 130.31	458	25	WH025, WH091, WH112, WH114, WH115, WH116, WH117, WH136, WH137, WH138, WH139, WH140, WH141, WH142, WH145, WH146, WH147, WH148, WH187, WH188, WH192, WH194, WH196, WH198, WH200
Linkou	Wuying	N 48.24; E 129.25	547	25	LK6, LK10, LK11, LK12, LK13, LK14, LK15, LK16, LK17, LK18, LK19, LK24, LK25, LK26, LK27, LK79-1, LK79-4, LK79-5, LK79-9, LK79-11, LK79-13, LK79-33, LK79-35, LK79-36, LK79-37
Tieli	Langxiang	N 46.95; E 128.87	332	22	TL1006, TL1018, TL1024, TL1054, TL1068, TL1080, TL1090, TL1091, TL1102, TL1105, TL1112, TL1140, TL1149, TL1185, TL1194, TL1198, TL1204, TL1212, TL1270, TL1271, TL1298, TL1357
Sanchazi	Sanchazi	N 42.63; E 126.85	601	24	SCZ113, SCZ114, SCZ115, SCZ116, SCZ117, SCZ119, SCZ120, SCZ121, SCZ122, SCZ123, SCZ124, SCZ125, SCZ126, SCZ127, SCZ129, SCZ130, SCZ131, SCZ132, SCZ133, SCZ134, SCZ135, SCZ136, SCZ137, SCZ138

Spectrophotometer (Bio-DL, Shanghai, China.) after extraction. The concentration of each DNA sample was diluted to 10 ng/μL and stored at -20°C.

2.2 SSR primer selection and genotyping

A total 142 primer pairs from the published SSR primers of 7 species of Pinaceae (*Pinus taeda*, *Pinus albicaulis*, *Pinus dabeshanensis*, *Pinus armandii*, *Pinus koraiensis* and *Pinus massoniana*) were selected and synthesized by Sangon Biotech (Shanghai) Co., Ltd., Shanghai, China (Liewlaksaneeyanawin et al., 2004; Echt et al., 2011; Yu et al., 2012; Dou et al., 2013; Xiang et al., 2015; Zhouxian et al., 2015; Zhang et al., 2015a; Dong et al., 2016; Lea et al., 2018; Li et al., 2020). Ten samples of DNA were randomly selected for polymorphism screening of synthesized primers. A PCR system was performed on DNA Engine thermal cycler (Biometra, Ilmenau OT Langewiesen, Germany) in 20μl volumes containing 0.5 μM each of forward and reverse primers, 200 μM dNTP, 2.0 μL 10×buffer, 2 U Taq DNA polymerase (TransGen Biotech Co., Beijing, China), and around 10 ng DNA. The PCR program was as follows: 3 min at 94°C, 35 cycles of 30 s at 94°C, 30 s at *Tm* (Table 2), and 15 s at 72°; and a final extension at 72°C for 7 min.

The PCR products were then detected by 7% PAGE, and 11 SSR markers with good reproducibility and significant polymorphism were selected finally. Forward primer of each marker was labelled at the 5' end with fluorescent dye HEX, 6-FAM, ROX, or TAM. PCR was performed under light-protected conditions with the same reaction system as above. All PCR products were sent to Sangon Biotech (Shanghai) Co., Ltd., Shanghai, China for capillary electrophoresis genotyping by ABI 3730XL (Applied Biosystems, Foster City, CA, USA) and the identification genotype data were collected for subsequent analysis.

2.3 Data analysis

2.3.1 Analysis of marker polymorphism and identification power

The DNA polymorphism information was processed into a data matrix, and the data matrix was converted into various formats by DataFormater 2.7 for further analysis (Wenqiang et al., 2016). Genetic parameters such as number of alleles (*Na*), number of effective alleles (*Ne*), Shannon diversity index (*I*), observed heterozygosity (*Ho*) and expected heterozygosity (*He*) of each primer was calculated using Popgen 32 (Li et al., 2003), primer polymorphism information content (*PIC*) was calculated using PowerMarker V3.25 (Liu and Muse, 2005), and primer identity probabilities (*PI*) and random identity probabilities (*PIsibs*) were calculated using GenAIEX 6.51b2 (Peakall and Smouse, 2006). Significant deviations from both the Hardy-Weinberg equilibrium (*HWE*) and linkage disequilibrium (*LD*) between all pairs of SSR loci were identified by Genepop v4.2 (Raymond and Rousset, 1995).

2.3.2 Genetic structure and genetic diversity analysis

GenAIEX 6.51b2 was used to calculate the number of alleles (*Na*), Shannon diversity index (*I*), number of effective alleles (*Ne*), number of more than 5% alleles (*Na, F>5*), observed heterozygosity (*Ho*), expected heterozygosity (*He*), unbiased expected heterozygosity (*uHe*), F-fixed index (*F*), and number of private loci (*NPA*) for each population (Peakall and Smouse, 2006). MANOVA, principal coordinates analysis (PCoA) and generation of interpopulation genetic differentiation coefficient (*Fst*) and gene flow (*Nm*) matrices were performed using GenAIEX 6.51b2 to delineate genetic variation between and within populations (Peakall and Smouse, 2006). Genetic distance matrix of clones clustering maps was generated by NtSys

TABLE 2 SSR primer information of Korean pine.

Locus	Primer Sequence	Motif	Tm (°C)	Size (bp)	Fluorescent dye	Reference
p49	F:GAGATGAGCGAATCTGGG	(AAG)7	52	261	FAM	Zhang et al., 2015b
	R:TACAAGTTCCACCTACGG					
p70	F:CAACATCGCCAATGACTC	(CTCA)6	54	294	FAM	
	R:CCTACCTACGCTCTGCTC					
p72	F:TGGGTTACCACCTTTAGC	(GCT)6	52	193	HEX	
	R:CAATCAGAGTCTGGAGCA					
p79	F:CCACCGCCAAGTCCATTA	(CAA)7	55	190	HEX	
	R:GCTTTGTTAGCCGTCCAG					
p82	F:GGAAGATGAATCGCAAACC	(GCG)6	54	280	ROX	
	R:ACACCCGCCTGAAGAGCA					
EPD11	F:GTGGATGCAATGAAGAAAAACAT	(AGG)6	60	139	TAM	Xiang et al., 2015
	R:ACGAATTGCAAAACTGCATAACT					
NFPK-34	F:AACCCACAGAAAGCTGAGGA	(TAA)6	60	221	TAM	Li et al., 2020
	R:CACCCCTGAACAGAGAGGAG					
P6*	F:TCAAATTACCAGACAATAA	(TA)3(GT)15	55	125	FAM	Yu, 2012
	R:GAATTCGCCAATGAAATCA					
P45*	F:CTTACATTTTGCTGCTTTTC	(TG)16(AG)17	55	173	HEX	
	R:TTGTCAGTTTTAGGTTGGAT					
P51*	F:CCTAAGAGCAATGTAAAATG	(AG)15	55	204	TAM	
	R:AGCTTGACAACGACTAACT					
P52*	F:CCATCCTTCAAATTTTCCT	(AG)26	56	138	ROX	
	R:GCCATTCTTTCTACCACTT					

2.10e and used for constructing a neighbor-joining dendrogram in MEGA 11 (Tohme et al., 1996; Tamura et al., 2011). Neighbor-joining dendrogram between populations was also constructed in MEGA 11 based on Nei genetic distance (Tamura et al., 2011).

Based on the latitude and longitude of the source location of superior tree, the geographical distance between the source locations of superior tree was calculated by the following formula:

$$d = R \star^{\arccos} \left[\cos(Y1) \star^{\cos} (Y2) \star^{\cos} (X1 - X2) + \sin(Y1) \star^{\sin} (Y2)\right]$$

R is the radii of the earth (6371.0 km);

X1, X2, Y1, Y2 are two location coordinates radians;

Radians= coordinates * Π/180;

SPSS v19.0 software was used to detect the correlation between geographic distance and genetic distance among the 6 superior tree sources.

The genetic structure was investigated in software STRUCTURE v2.3.4 using an admixed model with 100,000 burn-ins followed by 100,000 iterations (Ravelombola et al., 2018). Markov Chain Monte Carlo iterations run 10 times of a number (K = 2-18) of genetically homogeneous clusters. The operation results were imported into the Structure Harvester website (https://taylor0.biology.ucla.edu/structureHarvester/) (Earl and vonHoldt, 2012), and the optimal K values were selected according to the method of Evanno et al. (Evanno et al., 2005).

2.3.3 Fingerprint mapping construction

The fingerprint map of 161 clones was generated by combining the 11 pairs of SSR primers obtained from screening, sorting them in order from smallest to largest according to the size of the target fragment. The clone genotypes were coded using letters and arranged in a certain order to obtain the clone gene code. The name, scientific name and location of the clone, the source of the superior tree and the fingerprint code were organized into a separate Excel and uploaded to the online platform (https://cli.im/) to obtain the corresponding QR code for each clone (Li et al., 2022).

3 Result

3.1 Analysis of SSR marker polymorphism and discriminatory ability

There were 55 combinations of loci in the whole population, of which 4 pairs (7.27%) had LD between loci combinations at the

significance level of P<0.001. The results of primer genetic analysis (Table 3) showed that a total of 75 alleles were detected at the 11 SSR loci, of which 33.004 were effective alleles, with the mean value of major allele frequency was 0.622. 186 genotypes were identified, with an average of 16.909 genotypes per marker, and the observed heterozygosity and expected heterozygosity on average were 0.451 and 0.514, respectively. The mean values of Shannon diversity index and Nei diversity index were 1.094 and 0.512, respectively, indicating the high genetic diversity among clones. The polymorphism information content (*PIC*) of 11 markers ranged from 0.155 to 0.855, among which 10 markers showed moderate or high polymorphism relatively. All markers can effectively analyze the genetic structure and genetic diversity of Korean pine clones.

Two key statistical values, PI and PIsibs, were calculated to assess the ability to identify 11 markers for Korean pine clones (Table 3). PI for each molecular marker ranged from 0.028-0.593 with a mean value of 0.393. PIsibs is often defined as the upper PI limit, and the PIsibs of the 11 SSR markers ranged from 0.320-0.773 with a mean value of 0.565. The cumulative probability of identity of markers according to the obtained data (Figure 1), PI tended to 0 when the number of marker combinations is 7 and PIsibs tended to 0 when the number of marker combinations is 11. Assuming that all marker loci are independent of each other, the probability of two random Korean pine clones having the exact same multi-locus genotype combination among all 11 molecular markers is estimated to be 3.1×10^{-8}, and the combined PIsibs was 1.14×10^{-3}. 161 Korean pine clones could be considered to be completely distinguished by the 11 SSR markers. The above results prove that the combination of these markers not only had high polymorphism but also showed a strong potential for fingerprint recognition.

3.2 Analysis of genetic structure and genetic diversity

3.2.1 Analysis of genetic variation among populations

MANOVA was performed to determine variate characteristics of the 7 populations, and the results showed that (Table 4): population genetic differentiation coefficient (*Fst*) was 0.044 (P< 0.001), indicating a low level of genetic differentiation among populations. Genetic variation existed mainly within populations, accounting for 98% of the total variation, and the incidence of genetic variation among populations was only 2%. All of which indicated that there were extensive exchanges of genetic resources within each population. The level of genetic differentiation between populations was low, while the genetic variation within populations was much higher than that between populations. The inbreeding coefficient (*Fis*) was 0.078 (*Fis* > 0), indicating the presence of homozygous excess and the presence of interpopulation inbreeding.

Fst and *Nm* between two populations were calculated for seven populations to reveal genetic differences and gene flow among different populations of Korean pine clones. The results showed that (Figure 2): the *Fst* ranged from 0.012-0.047 with an average of 0.025, and the *Nm* ranged from 5.013-19.750 with an average of 11.036 among different populations, indicating that the genetic differentiation range among populations was small and there was a high frequency of genetic exchange. The highest *Fst* and the lowest *Nm* were found between the Lushuihe and Weihe, which may be due to the long geographical distance between Lushuihe and Hebei, the superior tree source of these two populations (Figure 3).

TABLE 3 Genetic diversity parameters of 11 SSR marker.

Locus	MAF	Na	Ne	N	Ho	He	Shannon	Nei	PIC	HWE	PI	PIsibs
p49	0.665	4	1.831	6	0.503	0.455	0.698	0.454	0.362	NS	0.418	0.646
p70	0.693	4	1.901	7	0.391	0.476	0.845	0.474	0.428	NS	0.319	0.590
p72	0.845	3	1.374	4	0.273	0.273	0.524	0.272	0.251	NS	0.545	0.748
p79	0.736	5	1.680	6	0.404	0.406	0.705	0.405	0.347	***	0.331	0.599
p82	0.832	4	1.400	5	0.186	0.287	0.530	0.286	0.256	***	0.533	0.737
EPD11	0.627	4	2.036	8	0.528	0.510	0.860	0.509	0.431	NS	0.321	0.580
NFPK-34	0.907	2	1.203	3	0.124	0.170	0.310	0.169	0.155	NS	0.593	0.773
P6*	0.640	9	2.297	22	0.534	0.566	1.287	0.565	0.543	NS	0.203	0.513
P45*	0.245	16	7.362	48	0.708	0.867	2.256	0.864	0.851	***	0.028	0.320
P51*	0.252	13	7.562	48	0.745	0.871	2.239	0.868	0.855	***	0.029	0.322
P52*	0.401	11	4.358	29	0.559	0.773	1.781	0.771	0.745	***	0.076	0.383
Mean	0.622	6.818	3.000	16.909	0.451	0.514	1.094	0.512	0.475		0.309	0.565
Total	–	75	33.004	186	–	–	–	–	–		3.1×10^{-8}	1.14×10^{-3}

***Denotes Significant departure from Hardy-Weinberg equilibrium at P<0.001. NS denotes meet Hardy-Weinberg equilibtium.

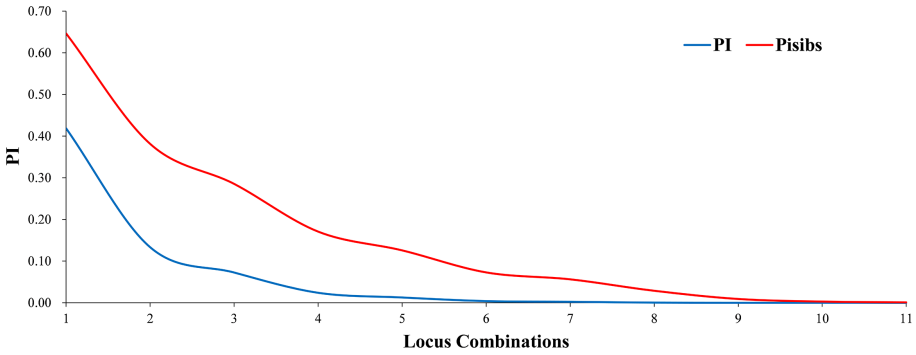


FIGURE 1
Identification ability of SSR markers in Korean pine clones.

3.2.2 Analysis of genetic diversity within populations

To assess genetic diversity and genetic differentiation of these 7 populations, genetic diversity analysis was performed and results showed that (Table 5): the level of genetic differentiation within 7 populations did not vary significantly, with Tieli population having the highest genetic diversity and the highest number of alleles, Shannon diversity index, observed heterozygosity at 55, 1.087, 0.479 respectively. the lowest Shannon diversity index and observed heterozygosity was in Lushuihe population with 0.915, 0.473 respectively.

The fixation index (*F*) ranged from -0.115 (Weihe) to 0.128 (Tieli), with an average of 0.061. F>0 indicates heterozygote deficiency, over-purity and inbreeding in Korean pine populations. Overall, the Tieli population showed high genetic diversity, while the Weihe population showed relatively lower genetic diversity, and no inbreeding was detected in this population.

The results of principal coordinate analysis (PCoA) of Korean pine clones from 7 populations showed that Coordinates 1 explained 9.93% of the variation and Coordinates 2 explained 7.45% of the variation, indicating that each of the above molecular markers has a high degree of independence. There is a high degree of distribution overlap among populations in the figure, with the Lushuihe population having an extensive distribution and some clones showing relative independence, while the other populations are relatively clustered, with the Linkou population being more dispersed. There is some genetic divergence between the Weihe population and the Tieli population, and the distribution range of Weihe is the smallest, indicating that the genetic diversity of clones within the Weihe population is low (Figure 4), which is similar to the results of the Shannon diversity index in Table 5.

3.2.3 Analysis of cluster and genetic structure

The results of cluster analysis among populations by the Nei genetic distance matrix showed (Figure 5A): The genetic distance among populations was small and the level of genetic differentiation was low, which was consistent with the results of MANOVA. The genetic distance between Bohai and Hegang, Lushuihe and Linkou was similar respectively, but the genetic distance of Weihe was far from Sanchazi. Lushuihe and Sanchazi were more independent, which was similar to the results of PCoA. However, it is worth noting that Hegang and Linkou have the same source location of superior tree, but they do not have the closest genetic relationship with each other. The correlation analysis between Nei genetic distance and geographic distance revealed that the Person coefficient was 0.075 (P=0.704), indicating the insignificant correlation between genetic distance and geographic distance of their superior tree sources.

The genetic distances of 161 clones were calculated by NTsys 2.10e software, and the results of clustering using MEGA showed that (Figure 5B): the clones from different sources were not clearly separated from each other, and the clones in each cluster did not come from the same location or the same superior source. The clones from different places were dispersed in each cluster. Clustering results did not correlate significantly with the location of the clones. The above results indicate that there is a high degree of gene exchange among populations and little genetic differentiation among populations. However, clones from Changbai Mountain are highly distributed on the left side of the clustering map, while clones from Xiaoxinganling are highly distributed on the right side and the lower part of the clustering map in general. Similar to the results of the principal coordinate analysis, although the populations were not clearly divided, the clones of different populations had

TABLE 4 MANOVA for the population of Korean pine clones.

Source	DF	SS	MS	Variance component	Variance component/%	Fit	Fis	Fst
Among Pops	6	1244.936	207.489	2.767	2			
Within Pops	154	22172.480	143.977	143.977	98			
Total	160	23417.416		146.745	100	0.117 ***	0.078 ***	0.044 ***

***Denotes significant differences at P<0.001.

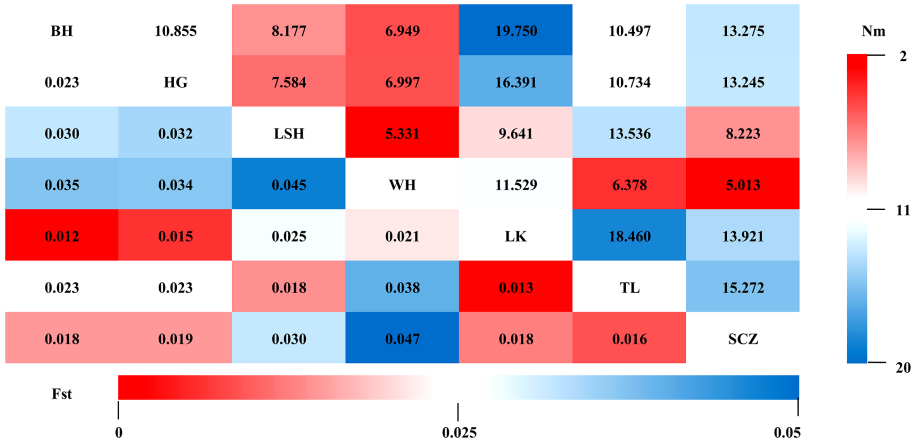


FIGURE 2
Genetic differentiation coefficients (lower left) and gene flow (upper right) between populations. (BH, Bohai; HG, Hegang; LSH, Lushuihe; WH, Weihe; LK, Linkou; TL, Tieli; SCZ, Sanchazi).

corresponding distribution ranges. For example, clones from Weihe had a small and relatively concentrated distribution range, which was consistent with the results of analysis of population diversity and the principal coordinate analysis, indicating that the genetic relationships among populations were similar and reflecting the degree of genetic differentiation within populations.

Structure analysis was performed on all the reference materials. ΔK had a maximum value when K=4 and 7 in K=2-18 (Figures 6A, B), indicating that the 161 clones could be divided into 4 classes or 7 classes. The populations were not clearly differentiated and no individuals had 100% population affiliation in both cases (Figures 6C, D). However, the Lushuihe partial clones had a

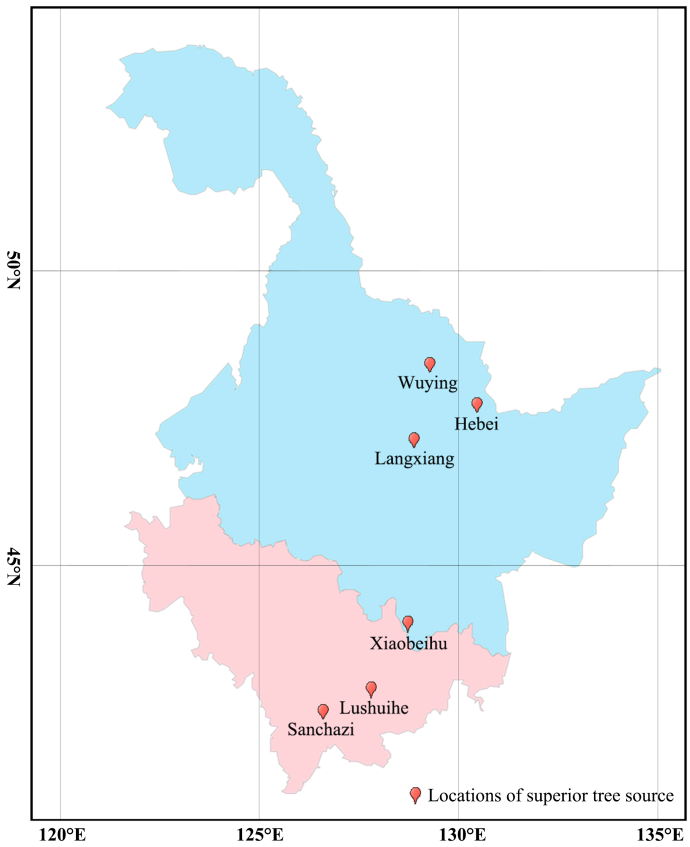


FIGURE 3
Geographical distribution of superior tree source of Korean pine populations. (Light blue indicates Heilongjiang Province and light purple indicates Jilin Province. The geographic conditions for each site can be found in Table 1).

TABLE 5 Genetic parameters of 7 Korean pine populations.

pop	Na	Ne	Na(F≥5%)	NPA	Shannon	Ho	He	uHe	F
Bohai	52	26.996	38.000	1	0.941	0.444	0.460	0.473	0.035
Hegang	52	30.149	38.000	0	1.002	0.448	0.505	0.515	0.083
Lushuihe	52	29.745	34.000	2	1.009	0.455	0.518	0.531	0.106
Weihe	53	31.431	36.000	0	0.915	0.473	0.437	0.445	-0.115
Linkou	52	31.084	39.000	1	1.001	0.425	0.488	0.498	0.119
Tieli	55	33.212	39.000	3	1.087	0.479	0.536	0.548	0.128
Sanchazi	53	30.790	38.000	2	1.012	0.436	0.502	0.512	0.074
Mean	52.714	30.487	37.429	1.286	0.995	0.451	0.492	0.503	0.061

significantly high probability of occurrence in a certain population, indicating that the Lushuihe population partial clones had relative genetic independence, which was similar to the results of PCoA.

3.3 Fingerprint mapping construction

Based on the genotyping data detected by 11 SSR markers, multiple locus matching analysis was performed in GenAlex 6.51 for 161 Korean pine clones. There is no identical genotype detected in two varieties, indicating that each of the 161 clones had its own unique SSR multi-locus genotype combination. The molecular markers were sorted according to the order of the target fragments from smallest to largest, and each marker consisted of two alleles. The molecular fingerprints of all 161 clones were generated according to the blocks with different color marking the different genotypes under each marker (Figure 7), with each color representing a variant locus information and each clone having a unique color block combination.

The genotyping data of each marker are indicated by letters respectively, and sorted in the order of amplified fragments from smallest to largest, and each clone gets its corresponding 22-digit letter code (Supplementary Table S1).

The name, location, source and genotyping data of each clone were uploaded to the QR code generation platform (https://cli.im/batch) to generate a unique QR code for each clone, which can be scanned to obtain specific information of the clone (Supplementary Figure S1).

4 Discussion

The genetic diversity of a population determines whether a population can adapt to a complex environment, and the higher the genetic diversity, the more adaptable the population is to different environments and the more resistant it is to shocks arising from environmental changes (Wachowiak et al., 2011). In order to develop a reasonable and effective breeding strategy, accelerate the process of genetic improvements of Korean pine, it is important to analyze the genetic diversity of Korean pine clone resources and evaluate the genetic structure of seed orchards by using SSR molecular marker. SSR markers have the advantages of codominance, stable amplification and good repeatability, which is a common method for genetic diversity analysis (Hao et al., 2017); at the same time, SSR molecular markers have strong specificity, clear bands and accurate data, which is suitable for the construction of fingerprint profile for a large number of resources (Park et al., 2009). Initially, we screened 11 Korean pine SSR loci, and the average values of *Na* and *He* for the 11 loci were 6.818, 0.514. *PIC* is an important parameter for expressing the degree of genetic diversity among plants, and its evaluation is beneficial to the establishment of plant gene pools and the acceleration of the breeding process (Avval, 2017). The average *PIC* of the SSR loci screened in this study was 0.475, showing moderate polymorphism (Botstein et al., 1980). Therefore, it is suitable for the genetic diversity evaluation of Korean pine breeding resources. The LD between 4 pairs of loci reached

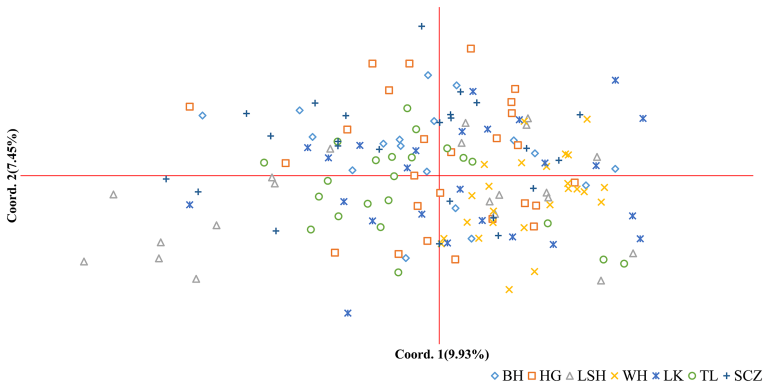


FIGURE 4
Principal coordinates analysis (PCoA) of Korean pine clones. (BH, Bohai; HG, Hegang; LSH, Lushuihe; WH, Weihe; LK, Linkou; TL, Tieli; SCZ, Sanchazi).

10.3389/fpls.2022.1079571

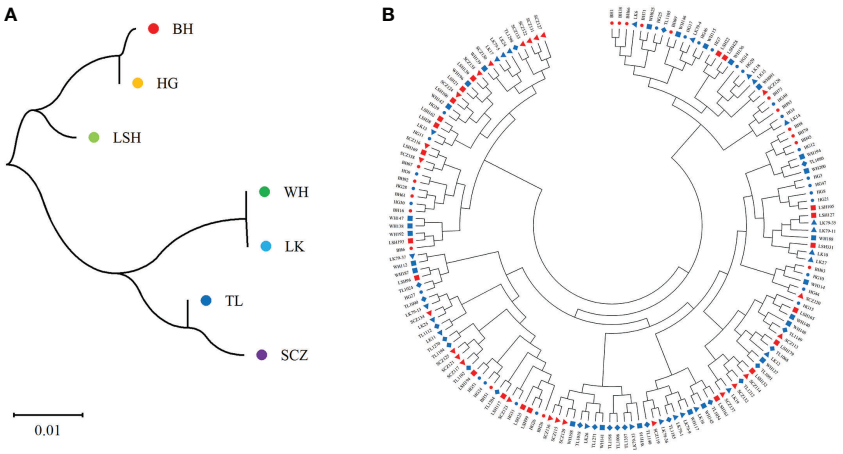


FIGURE 5
Neighbor-joining tree of populations and clones. **(A)** Neighbor-joining tree of 7 populations. **(B)** Neighbor-joining tree of 161Korean pine clons. (BH, Bohai; HG, Hegang; LSH, Lushuihe; WH, Weihe; LK, Linkou; TL, Tieli; SCZ, Sanchazi).

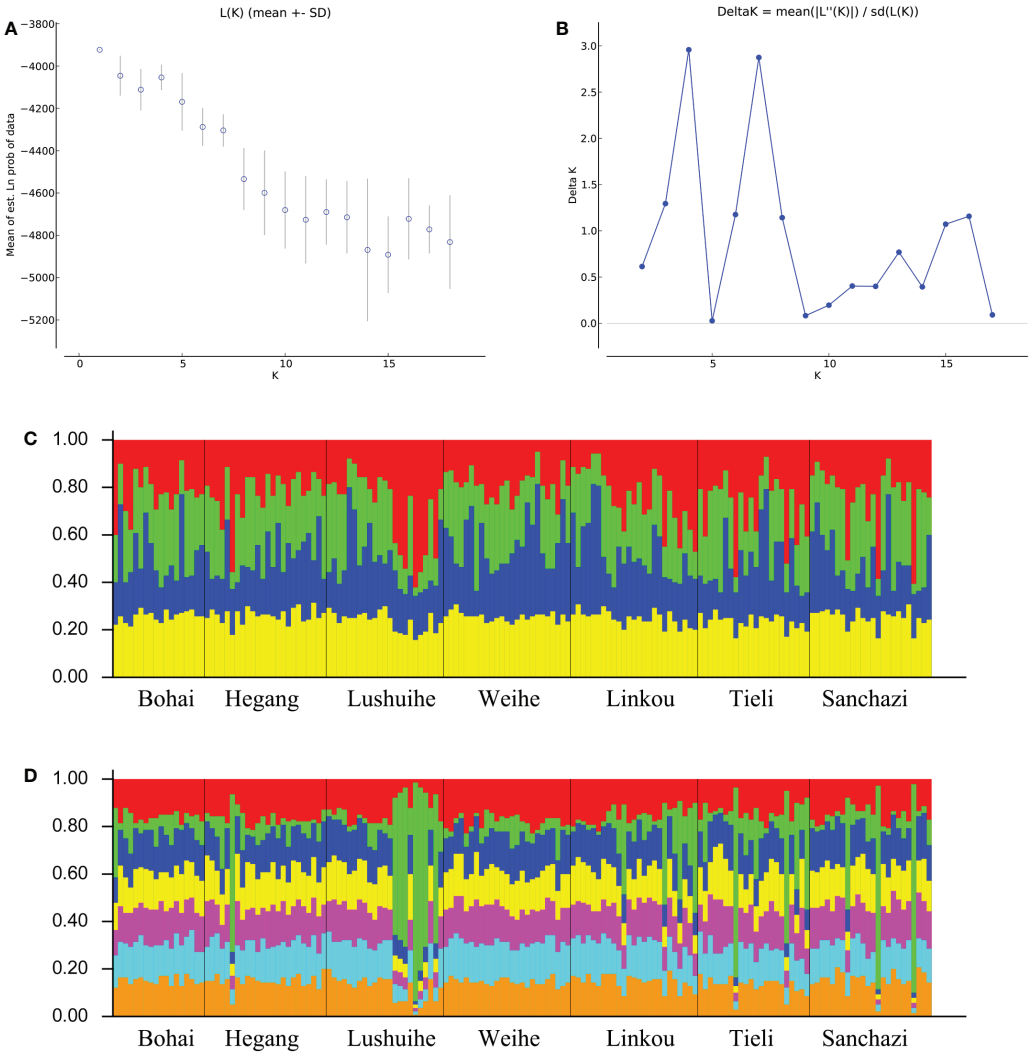


FIGURE 6
STRUCTURE analysis of Korean pine population. **(A)** Calculation of population structure using Mean LnP (K). **(B)** Relations between the optional number of cluster K and Delta K**(C)** Genetic structure map of 7 populations of Korean pine based on STRUCTURE analysis (K = 4). **(D)** Genetic structure map of 7 populations of Korean pine based on STRUCTURE analysis (K = 7).

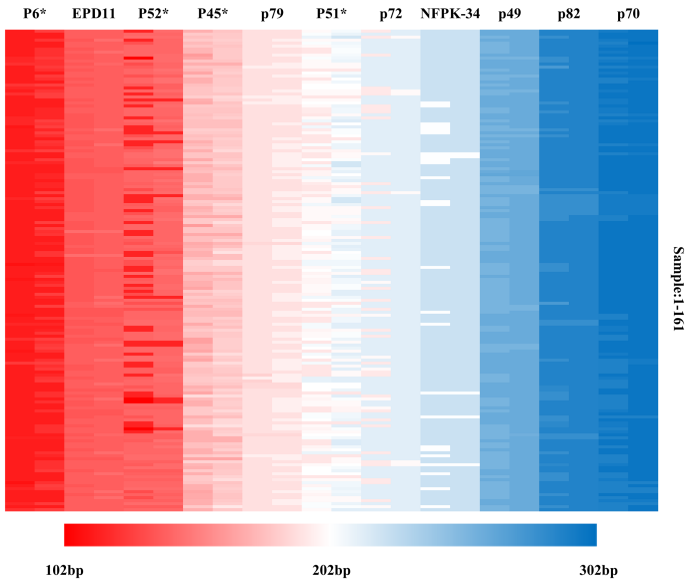


FIGURE 7
Molecular fingerprinting of Korean pine clones. (The different color blocks represent the corresponding allele fragment sizes).

a significant level of P<0.001, but was not concentrated at one locus, and the results of PCoA showed that the molecular markers were highly independent, indicating that the screened loci were evenly distributed in the Korean pine genome and relatively independent in the process of transmission from generation to generation.

In order to elucidate the genetic variation among Korean pine populations, molecular variation analysis was conducted. The results showed that the genetic variation of Korean pine mainly originated from inter-individuals, accounting for 98%, and interpopulation variation accounted for only 2%. This indicates that the genetic differentiation within populations is much greater than between populations, which is consistent with the results of Feng et al. (Feng et al., 2006). The result is consistent with higher genetic diversity within populations and higher gene flow between populations. Therefore, we should pay attention to the selection of individuals within the population when the Korean pine population with high genetic diversity was constructed in the later stage, which is beneficial to the genetic improvement of Korean pine. The genetic diversity analysis of 7 populations revealed the differences in the level of genetic diversity among different populations, Tieli has the highest level of genetic diversity (*I*=1.087), the genetic diversity of Weihe population was low (I=0.915). Nevertheless, Weihe population is the only one with a fixed index (*F*) less than 0, indicating that the genetic diversity of this population is low, but there is no heterozygosity deficiency or inbreeding. Heterozygosity is often used to measure the degree of genetic variation and can provide useful information for the conservation of species (Schmidt et al., 2021). The results of this study based on SSR molecular markers showed that the overall *He* and *Ho* of 7 populations were 0.514 and 0.451. From a biological point of view, Korean pine is a monoecious, cross-pollinated plant that can generate new genotypes through genetic recombination, which is probably the main reason why Korean pine populations maintain a high genetic diversity. The *Ho* is smaller than *He* among these populations, except for the Weihe population, which indicating the presence of heterozygote deficiency, this may be due to inbreeding, non-random mating or disruption of population structure (Liao et al., 2019). Therefore, further analysis for the reason of heterozygote deficiency is necessary in future studies.

Genetic structure reveals the distribution patterns of genetic diversity between and within populations, reflects the adaptive potential of various species to their environment (Melo et al., 2014). Seven Korean pine populations in this study can be divided into 4 or 7 classes, and different populations are mismatched in classifications. Lushuihe population shows partial independence relatively, and the corresponding results were obtained by clustering results, which is consistent with the results of the principal coordinate analysis mentioned above. The results of interpopulation differentiation also show that the Lushuihe population has higher genetic differentiation and lower gene flow with other populations, which may be due to the relatively isolated population structure caused by the relatively unique geographical location of Lushuihe. Correlation analysis showed that there was no significant correlation between genetic distance and sources of superior tree's geographical distance of Korean pine populations, which was also previously reported in Feng et al. (2009).

Screening and identifying the core SSR primer combinations suitable for variety identification is the key to constructing DNA fingerprinting. It is required that the core SSR primer combinations screened and identified have good marker polymorphism, and secondly, it is required that as few markers as possible are used to distinguish as many germplasms as possible. Construction of fingerprint profiles of Korean pine clones provides an important basis for the identification of resources from the 7 seed orchards. The DNA fingerprint profile of Korean pine clones based on SSR primer combinations can be directly used to identify the authenticity of clones in the 7 seed orchards, solving the long-standing problem of Korean pine clone identification. It is important for the selection and breeding of Korean pine clone in these 7 seed orchards. The critical point to ponder, the established fingerprint panel or Korean pine

clonal identification was based on 7 seed orchards in northeast China. It does not cover the distribution range of the species which also can be found in Korea, Russia, Mongolia & Japan. Hence, this clonal identification tool developed will solely useful within China (limited to the resources from the 7 seed orchards).

In this study, 11 SSR markers were screened out, which could be used for the construction of fingerprints of Korean pine clones and the evaluation of genetic structure of the population. Genetic analysis of 7 populations of Korean pine using 11 SSR primers revealed the level of genetic diversity and genetic differentiation among and within populations. According to the genetic characteristics of Korean pine clone populations, the development of corresponding breeding strategies can maximize the breeding potential of Korean pine seed orchards and provide a scientific basis for the subsequent development and utilization of Korean pine germplasm. The DNA fingerprints of 161 Korean pine clones were constructed, which is an effective strategy for the identification of Korean pine clone, it will provide strong DNA evidence for identification of variety and superior seed validation.

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.

Author contributions

Conceptualization: PY and HZ. Methodology: XQ. Validation: KF. Resources: ZX. Writing-original draft preparation: PY. Writing-review and editing: LZ. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Innovation Project of State Key Laboratory of Tree Genetics and Breeding (Northeast Forestry University) "Research on the Evaluation of Korean Pine Germplasm Resources and Construction Technology of Breeding Population" (2022A02), the Fundamental Research Funds for the Central Universities (2572022AW13) and Heilongjiang Touyan Innovation Team Program.

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.2022.1079571/full#supplementary-material

References

Afeng, Z. (2005). *Establishment of fingerprinting for clones in pinus massoniana with microsatellite markers* (Nanjing (Jiangsu province: Nanjing Forestry University). master's thesis.

Agarwal, M., Shrivastava, N., and Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.* 27 (4), 617–631. doi: 10.1007/s00299-008-0507-z

An, Z., Xangxuan, W., and Jixiang, L. (1992). A study on the pruning technique for seed high yield of Pinus koraiensis in seed orchard. *Scientia Silvae Sinicae*. 38(04), 349–352.

Avval, S. E. (2017). Assessing polymorphism information content (PIC) using SSR molecular markers on local species of citrullus colocynthis. case study: Iran, sistan-balouchestan province. *J. Mol. Biol. Res.* 7 (1).

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 (3), 314–331.

Carletti, G., Cattivelli, L., Vietto, L., and Nervo, G. (2021). Multiallelic and multilocus simple sequence repeats (SSRs) to assess the genetic diversity of aSalixspp. germplasm collection. *J. Forestry Res.* 32 (1), 263–271. doi: 10.1007/s11676-019-00913-0

Dong, W.-L., Wang, R.-N., Yan, X.-H., Niu, C., Gong, L.-L., and Li, Z.-H. (2016). Characterization of polymorphic microsatellite markers in Pinus armandii (Pinaceae), an endemic conifer species to China. *Appl. Plant Sci.* 4 (10). doi: 10.3732/apps.1600072

Dou, J. J., Zhou, R. C., Tang, A. J., Ge, X. J., and Wu, W. (2013). Development and characterization of nine microsatellites for an endangered tree, pinus wangii (Pinaceae). *Appl. Plant Sci.* 1. doi: 10.3732/apps.1200134

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 (2), 359–361. doi: 10.1007/s12686-011-9548-7

Echt, C. S., Saha, S., Deemer, D. L., and Nelson, C. D. (2011). Microsatellite DNA in genomic survey sequences and UniGenes of loblolly pine. *Tree Genet. Genomes* 7 (4), 773–780. doi: 10.1007/s11295-011-0373-7

Eding, H., Crooijmans, R. P. M. A., Groenen, M. A. M., and Meuwissen, T. H. E. (2002). Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genetics selection Evol. GSE* 34 (5), 613–633. doi: 10.1186/1297-9686-34-5-613

Ellegren, H., and Galtier, N. (2016). Determinants of genetic diversity. *Nat. Rev. Genet.* 17 (7), 422–433. doi: 10.1038/nrg.2016.58

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

Feng, F., Chen, M., Zhang, D., Sui, X., and Han, S. (2009). Application of SRAP in the genetic diversity of pinus koraiensis of different provenances. *Afr. J. Biotechnol.* 8 (6), 1000–1008.

Feng, F., Han, S., and Wang, H. (2006). Genetic diversity and genetic differentiation of natural pinus koraiensis population. *J. Forestry Res.* 17 (1), 21–24. doi: 10.1007/s11676-006-0005-5

Feng, F., Wang, F., and Liu, T. (2004). The influence factors of the ISSR-PCR experiment system on pinus koraiensis sieb.et zucc. *Chin. Bull. Bot.* 21 (3), 326–331.

Feng, F. J., Zhao, D., Sun, X. Y., Han, S. J., and Xiao-Yana, X. U. (2010). Establishment and optimization of the SSR-PCR reaction system in pinus koraiensis sieb.et zucc. *Nonwood For. Res.*

Flint-Garcia, S. A. (2013). Genetics and consequences of crop domestication. *J. Agric. Food Chem.* 61 (35), 8267–8276. doi: 10.1021/jf305511d

Gao, Y., Tian, L., Liu, F., and Cao, Y. (2012). Using the SSR fluorescent labeling to establish SSR fingerprints for 92 cultivars in pyrus. *Acta Hortic. Sin.* 39 (8), 1437–1446.

Govindaraj, M., Vetriventhan, M., and Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genet. Res. Int.* 2015, 431487–431487. doi: 10.1155/2015/431487

Guan, C., Zhang, P., Hu, C., Chachar, S., Riaz, A., Wang, R., et al. (2019). Genetic diversity, germplasm identification and population structure of diospyros kaki thunb. from different geographic regions in China using SSR markers. *Scientia Hortic.* 251, 233–240. doi: 10.1016/j.scienta.2021.110064

Hao, L., Zhang, L., Zhang, G., Wang, Y., Han, S., and Bai, Y. (2017). Genetic diversity and population genetic structure of salix psammophila. *Acta Botanica Boreali-Occidentalia Sin.* 37 (8), 1507–1516.

Hughes, A. R., Inouye, B. D., Johnson, M. T. J., Underwood, N., and Vellend, M. (2008). Ecological consequences of genetic diversity. *Ecol. Lett.* 11 (6), 609–623. doi: 10.1111/j.1461-0248.2008.01179.x

Huili, W., Shuxue, Y., Li, G., Hailong, L., and Wei, L. (2022). Genetic diversity assessment and fingerprint construction of superior tree populations of pinus sylvestris var. mongolica. *J. OF GANSU Agric. Univ.* 57 (003), 057. doi: 10.13432/j.cnki.jgsau.2022.03.013

Ivetić, V., Devetaković, J., and Maksimović, Z. (2016). Initial height and diameter are equally related to survival and growth of hardwood seedlings in first year after field planting. *REFORESTA* 2), 6–21. doi: 10.21750/REFOR.2.02.17

Ivetic, V., Devetakovic, J., Nonic, M., Stankovic, D., and Sijacic-Nikolic, M. (2016). Genetic diversity and forest reproductive material - from seed source selection to planting. *Iforest-Biogeosciences Forestry* 9, 801–812. doi: 10.3832/ifor1577-009

Jin, Y., Ma, Y., Wang, S., Hu, X.-G., Huang, L.-S., Li, Y., et al. (2016). Genetic evaluation of the breeding population of a valuable reforestation conifer platycladus orientalis (Cupressaceae). *Sci. Rep.* 6. doi: 10.1038/srep34821

Jump, A. S., Marchant, R., and Penuelas, J. (2009). Environmental change and the option value of genetic diversity. *Trends Plant Sci.* 14 (1), 51–58. doi: 10.1016/j.tplants.2008.10.002

Lea, M. V., Syring, J., Jennings, T., Cronn, R., Bruederle, L. P., Neale, J. R., et al. (2018). Development of nuclear microsatellite loci for pinus albicaulis engelm. (Pinaceae), a conifer of conservation concern. *PloS One* 13 (10). doi: 10.1371/journal.pone.0205423

Ledig, F. T. (1992). Human impacts on genetic diversity in forest ecosystems. *Oikos* 63 (1), 87–108. doi: 10.2307/3545518

Liang, H., Liu, C., Liu, X., and Yang, M. (2005). Simple sequence repeat (SSR) analysis and identify of different cultivars in populus l. *J. Agric. Univ. Hebei* 28 (4), 27–31.

Liao, R., Luo, Y., Yisilam, G., Lu, R., Wang, Y., and Li, P. (2019). Development and characterization of SSR markers for sanguinaria canadensis based on genome skimming. *Appl. Plant Sci.* 7 (9). doi: 10.1002/aps3.11289

Liewlaksaneeyanawin, C., Ritland, C. E., El-Kassaby, Y. A., and Ritland, K. (2004). Single-copy, species-transferable microsatellite markers developed from loblolly pine ESTs. *Theor. Appl. Genet.* 109 (2), 361–369. doi: 10.1007/s00122-004-1635-7

Li, L., Lejing, L., Zhiyong, Z., Bo, L., and Jie, R. (2022). Construction of SSR fingerprint and genetic diversity analysis of 93 maple germplasm resources. *Mol. Plant Breed.* 20 (4), 1250–1263. doi: 10.13271/j.mpb.020.001250

Li, X., Liu, X., Wei, J., Li, Y., Tigabu, M., and Zhao, X. (2020). Development and transferability of EST-SSR markers forPinus koraiensisfrom cold-stressed transcriptome through illumina sequencing. *Genes* 11 (5). doi: 10.3390/genes11050500

Lim, T. K. (2012). *Edible medicinal and non medicinal plants* (Netherlands: Springer Netherlands).

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

Li, T., Yang, W., Song, L., Su, X., Yang, Z., and Guo, H. (2003). Exploring on the genetic polymorphism in haliotis discus hannai ino and h. diversicolor reeve by RAPD technique. *Oceanologia Limnologia Sin.* 34 (4), 444–449.

Longhai, H., Fenggang, P., Hongzhi, L., Xiaozhong, S., Lianjun, Y., and Hanguo, Z. (2021). Variation of cone-and-seed traits and clonal selection of pinus koraiensis. *J. Beihua University (Natural Science)* 22 (2), 6.

Lv, J., Li, C., Zhou, C., Chen, J., Li, F., Weng, Q., et al. (2020). Genetic diversity analysis of a breeding population of eucalyptus cloeziana f. muell. (Myrtaceae) and extraction of a core germplasm collection using microsatellite markers. *Ind. Crops Products* 145. doi: 10.1016/j.indcrop.2020.112157

Ma, L., Kong, D., Liu, H., Wang, S., Li, Y., and Pang, X. (2012). Construction of SSR fingerprint on 36 Chinese jujube cultivars. *Acta Hortic. Sin.* 39 (4), 647–654.

Markert, J. A., Champlin, D. M., Gutjahr-Gobell, R., Grear, J. S., Kuhn, A., McGreevy, T. J.Jr., et al. (2010). Population genetic diversity and fitness in multiple environments. *BMC Evolutionary Biol.* 10. doi: 10.1186/1471-2148-10-205

Melo, A. T. D., Coelho, A. S. G., Pereira, M. F., Blanco, A. J. V., and Franceschinelli, E. V. (2014). High genetic diversity and strong spatial genetic structure in cabralea canjerana (Vell.) mart. (Meliaceae): implications to Brazilian Atlantic forest tree conservation. *Natureza Conservacao* 12 (2), 129–133. doi: 10.1016/j.ncon.2014.08.001

Nergiz, C., and Donmez, I. (2004). Chemical composition and nutritive value of pinus pinea l. seeds. *Food Chem.* 86 (3), 365–368. doi: 10.1016/j.foodchem.2003.09.009

Nevo, E. (1988). Genetic diversity in nature: Patterns and theory. *Evolutionary Biol.* 23, 217–246. doi: 10.1007/978-1-4613-1043-3_6

Nn, A., Iss, A., As, A., Sd, A., Mkp, B., La, A., et al. (2020). Population genetic structure and diversity analysis in economically important pandanus odorifer (Forssk.) kuntze accessions employing ISSR and SSR markers. *Ind. Crops Products* 143. doi: 10.1016/j.indcrop.2019.111894

Park, Y.-J., Lee, J. K., and Kim, N.-S. (2009). Simple sequence repeat polymorphisms (SSRPs) for evaluation of molecular diversity and germplasm classification of minor crops. *Molecules* 14 (11), 4546–4569. doi: 10.3390/molecules14114546

Pauls, S. U., Nowak, C., Balint, M., and Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Mol. Ecol.* 22 (4), 925–946. doi: 10.1111/mec.12152

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: Genetic analysis in excel. population genetic software for teaching and research. *Mol. Ecol. Notes* 6 (1), 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Pingyu, Y., Peng, W., Weiman, Y., and Hanguo, Z. (2020). Analysis of seeding characters of Korean pine seed orchard and selection of excellent clones. *For. Eng.* 36 (6), 11. doi: 10.16270/j.cnki.slgc.2020.06.003

Qianping, T. (2020). Genetic diversity analysis on indivitual of pinus koraiensis in seed orchard based on ISSR-PCR. *For. Sci. Technol.* 45 (2), 4. doi: 10.19750/j.cnki.1001-9499.2020.02.005

Ravelombola, W., Shi, A., Weng, Y., Mou, B., Motes, D., Clark, J., et al. (2018). Association analysis of salt tolerance in cowpea (Vigna unguiculata (L.) walp) at germination and seedling stages. *Theor. Appl. Genet.* 131 (1), 79–91. doi: 10.1007/s00122-017-2987-0

Raymond, M., and Rousset, F. (1995). GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicism. *J. Heredity* 68 (3), 248–249.

Schmidt, T. L., Jasper, M.-E., Weeks, A. R., and Hoffmann, A. A. (2021). Unbiased population heterozygosity estimates from genome-wide sequence data. *Methods Ecol. Evol.* 12 (10), 1888–1898. doi: 10.1111/2041-210x.13659

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* 28 (10), 2731–2739. doi: 10.1093/molbev/msr121

Tohme, J., Gonzalez, D. O., Beebe, S., and Duque, M. C. (1996). AFLP analysis of gene pools of a wild bean core collection. *Crop Sci.* 36. doi: 10.2135/cropsci1996.0011183X003600050048x

Tong, Y., Tang, Y., Chen, H., Zhang, T., Zuo, J., Wu, J., et al. (2019). Phenotypic diversity of pinus koraiensis populations in a seed orchard. *Acta Ecologica Sin.* 39 (17), 6341–6348.

Wachowiak, W., Salmela, M. J., Ennos, R. A., Iason, G., and Cavers, S. (2011). High genetic diversity at the extreme range edge: Nucleotide variation at nuclear loci in scots pine (Pinus sylvestris l.) in Scotland. *Heredity* 106 (5), 775–787. doi: 10.1038/hdy.2010.118

Weihuai, W., Yutong, C., Qianshun, D., Dan, H., Zhixin, L., Hanguo, Z., et al. (2019). Analysis of characters and nutritional components of pinus koraiensis seeds on clones and superior trees. *For. Eng.* 35 (2), 11. doi: 10.16270/j.cnki.slgc.2019.02.002

Wenqiang, F., Hongmei, G., Xin, S., Aiguo, Y., Zhongfeng, Z., and Min, R. (2016). DataFormater, a software for SSR data formatting to develop population genetics analysis. *Mol. Plant Breed.* 1), 6. doi: 10.13271/j.mpb.014.000265

Wheeler, N. C., and Jech, K. S. (1992). The use of electrophoretic markers in seed orchard research. *New Forests* 6 (1-4), 311–328. doi: 10.1007/BF00120650

Wolff, R. L., Pédrono, F., Pasquier, E., and Marpeau, A. M. (2000). General characteristics of pinus spp. seed fatty acid compositions, and importance of Δ5-olefinic acids in the taxonomy and phylogeny of the genus. *Lipids* 35 (1), 1–22. doi: 10.1007/s11745-000-0489-y

Xiang, X., Zhang, Z., Wang, Z., Zhang, X., and Wu, G. (2015). Transcriptome sequencing and development of EST-SSR markers in pinus dabeshanensis, an endangered conifer endemic to China. *Mol. Breed.* 35 (8). doi: 10.1007/s11032-015-0351-0

Yoon, T. H., Im, K. J., Koh, E. T., and Ju, J. S. (1989). Fatty acid compositions of pinus koraiensis seed. *Nutr. Res.* 9 (3), 357–361. doi: 10.1016/S0271-5317(89)80079-X

Yu, J.-H., Chen, C.-M., Tang, Z.-H., Yuan, S.-S., Wang, C.-J., and Zu, Y.-G. (2012). Isolation and characterization of 13 novel polymorphic microsatellite markers for *Pinus koraiensis* (Pinaceae). *Am. J. Bot.* 99 (10), E421–E424. doi: 10.3732/ajb.1200145

Zadernowski, R., Naczk, M., and Czaplicki, S. (2009). Chemical composition of pinus sibirica nut oils. *Eur. J. Lipid Sci. Technol.* 111 (7), 698–704. doi: 10.1002/ejlt.200800221

Zhang, Z., Zhang, H., Mo, C., and Zhang, L. (2015a). Transcriptome sequencing analysis and development of EST-SSR markers for pinus koraiensis. *Scientia Silvae Sinicae* 51 (8), 114–120.

Zhang, Q., Zhang, X., Yang, Y., Xu, L., Feng, J., Wang, J., et al. (2022). Genetic diversity of juglans mandshurica populations in northeast China based on SSR markers. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.931578

Zhang, Z., Zhang, H., Zhou, Y., Liu, L., Yu, H., Wang, X., et al. (2015b). Variation of seed characters in Korean pine (Pinus koraiensis) multi-clonal populations. *J. Beijing Forestry Univ.* 37 (2), 067–068.

Zhao, D., Zhang, D. D., Xin, S., and Feng, F. J. (2010). Establishment and optimization of the SRAP-PCR reaction system on pinus koraiensis sieb.et zucc. *Res. Explor. Laboratory.*

Zhouxian, N., Tiandao, B., Heng, C., Shufen, C., and Lian, X. (2015). The transferability of pinus massoniana SSR in other pinus species. *Mol. Plant Breed.* 13 (12), 7. doi: 10.13271/j.mpb.013.002811

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 06 February 2023
DOI 10.3389/fpls.2023.1112748

OPEN ACCESS

EDITED BY
Yuri Shavrukov,
Flinders University, Australia

REVIEWED BY
Reetika Mahajan,
Sher-e-Kashmir University of Agricultural Sciences and Technology, India
Shahbaz Khan,
National Agricultural Research Center, Pakistan

*CORRESPONDENCE
Ruixi Han
wudifeixue007@163.com
Zhenjiang Xu
zhenjiangxu521@scau.edu.cn

†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 30 November 2022
ACCEPTED 16 January 2023
PUBLISHED 06 February 2023

CITATION
Yin J, Zhao H, Wu X, Ma Y, Zhang J, Li Y, Shao G, Chen H, Han R and Xu Z (2023) SSR marker based analysis for identification and of genetic diversity of non-heading Chinese cabbage varieties.
Front. Plant Sci. 14:1112748.
doi: 10.3389/fpls.2023.1112748

SSR marker based analysis for identification and of genetic diversity of non-heading Chinese cabbage varieties

Jiwei Yin[1†], Hong Zhao[2†], Xingting Wu[3], Yingxue Ma[3], Jingli Zhang[2], Ying Li[4], Guirong Shao[5], Hairong Chen[2], Ruixi Han[3*] and Zhenjiang Xu[1*]

[1]College of Agriculture, South China Agricultural University, Guangzhou, China, [2]Institute for Agri-food Standards and Testing Technology, Shanghai Academy of Agricultural Sciences, Shanghai, China, [3]Development Center of Science and Technology, Ministry of Agriculture and Rural Affairs, Beijing, China, [4]State Key Laboratory of Crop Genetics and Germplasm Enhancement, College of Horticulture, Nanjing Agricultural University, Nanjing, China, [5]Research and development center, Fujian Jinpin Agricultural Technology Co., Ltd, Fuzhou, China

As a widely cultivated vegetable in China and Southeast Asia, the breeding of non-heading Chinese cabbage (*Brassica campestris* ssp. *chinensis* Makino) is widespread; more than 400 varieties have been granted new plant variety rights (PVRs) in China. Distinctness is one of the key requirements for the granting of PVRs, and molecular markers are widely used as a robust supplementary method for similar variety selection in the distinctness test. Although many genome-wide molecular markers have been developed, they have not all been well used in variety identification and tests of distinctness of non-heading Chinese cabbage. In this study, by using 423 non-heading Chinese cabbage varieties collected from different regions of China, 287 simple sequence repeat (SSR) markers were screened for polymorphisms, and 23 core markers were finally selected. The polymorphic information content (PIC) values of the 23 SSR markers ranged from 0.555 to 0.911, with an average of 0.693, and the average number of alleles per marker was 13.65. Using these 23 SSR markers, 418 out of 423 varieties could be distinguished, with a discrimination rate of 99.994%. Field tests indicated that those undistinguished varieties were very similar and could be further distinguished by a few morphological characteristics. According to the clustering results, the 423 varieties could be divided into three groups: pak-choi, caitai, and tacai. The similarity coefficient between the SSR markers and morphological characteristics was moderate (0.53), and the efficiency of variety identification was significantly improved by using a combination of SSR markers and morphological characteristics.

KEYWORDS

non-heading Chinese cabbage, SSR, variety identification, DUS test, genetic diversity

1 Introduction

Non-heading Chinese cabbage (*Brassica campestris* ssp. *chinensis* Makino) is a subspecies of *Brassica* in the family Brassicaceae. It is usually diploid (2n = 20, AA) and comprises five types: pak-choi (var. *communis* Tesn et Lee), caitai (var. *tsai-tai* Hort), tacai (var. *rosularism* Tesn et Lee), taicai (var. *tai-tsai* Hort), and duotoucai (var. *multiceps* Hort) (Hou and Song, 2012). Originating in China, it has a long history of cultivation, with its leaves being the main product (Hou et al., 2020). Given its strong adaptability, short growth cycle, and rich nutritional value, the non-heading Chinese cabbage is widely planted not only in China, but also in Southeast Asia, Europe, and America, and is gradually becoming a global vegetable. Non-heading Chinese cabbage breeding is widespread in China. As of August 2022, the number of applications for plant variety rights (PVRs) of non-heading Chinese cabbage in China had reached 436. However, owing to the lack of outstanding inbred lines and germplasm resources, the genetic background of newly developed varieties of non-heading Chinese cabbage is becoming narrower, and variety identification is becoming more and more difficult.

A distinctness, uniformity, and stability (DUS) test is the key technical support for the granting of PVRs, in which the distinctness test is the key step. To assess distinctness, the candidate variety needs to be compared with any other commonly known varieties. To ensure the effectiveness and accuracy of the distinctness assessment, the construction of the database of commonly known varieties is very urgent and necessary. The effectiveness and accuracy of any distinctness assessment relies on the existence of a comprehensive and accurate database of commonly known varieties. The currently available database is based on the morphological characterization of commonly known varieties, which, although accurate and scientifically robust, also has several limitations, being slow, expensive, resource intensive, and time-consuming (Liu et al., 2012). In addition, as morphological characteristics are easily affected by environmental factors, such as temperature, light, and fertilizer application, and data collected in different ecological places or in different seasons may be quite different, which may cause errors when screening for similar varieties using the distinctness test. DNA molecular marker technology can directly detect differences on a DNA level among varieties; this technology is not easily affected by environmental conditions, does not require field planting, and is fast and efficient. It has been widely used in variety identification and is recommended as a supplementary method to construct a variety database for variety management, especially for screening for similar varieties using the distinctness test developed by the UPOV (International Union for the Protection of New Varieties of Plants). In contrast to other molecular markers, simple sequence repeat (SSR) markers have the advantages of clear loci, simple technology, and reliable detection results, and are recommended as one of the preferred markers for plant variety identification and database construction by UPOV (Zhou et al., 2020). SSR markers have been widely used in identifying Brassicaceae Burnett vegetables, such as Chinese cabbage, *Brassica* juncea, broccoli, and cauliflower (Zhan et al., 2014; Yan et al., 2021; Chu et al., 2020), and are also used in the identification and genetic diversity assessment of non-heading Chinese cabbage (Wang et al., 2008; Liu et al., 2014; Yu et al., 2014; Liu et al., 2021). However, in previous studies in non-heading Chinese cabbage, the SSR markers selected were comparatively low in polymorphism and could not meet the needs of large-scale variety identification (Li, 2010; Liu et al., 2014). The varieties that could be identified were usually limited to one or a few types, and did not cover all five types of non-heading Chinese cabbage; in addition, SSR data were mainly obtained by polyacrylamide gel electrophoresis, which was not conducive to genotyping and data-sharing. Therefore, it is necessary to establish a high-throughput SSR molecular identification system with a strong discriminatory ability that covers various types of non-heading Chinese cabbage, and which may provide a robust technical support for screening similar varieties using distinctness tests, identification of variety authenticity, and protection of PVRs.

In this study, by using non-heading Chinese cabbage varieties covering all the five types from all the main production areas in China, we tried to select a core set of SSR markers with high levels of polymorphism and strong discriminatory ability suitable for both polyacrylamide gel electrophoresis and capillary electrophoresis platforms. Based on the core SSR markers, an SSR fingerprint database could be constructed to provide a powerful support for similar variety screening of distinctness tests and variety identification of non-heading Chinese cabbage varieties.

2 Materials and methods

2.1 Plant materials and DNA extraction

Non-heading Chinese cabbage has rich morphological diversity and exhibits significant differences among variety types (Figure 1). A total of 423 non-heading Chinese cabbage varieties covering five subspecies, pak-choi, caitai, tacai, taicai, and duotoucai (Table S1), were collected in this study, among which two varieties were from northeast China, 40 were from north and central China, 36 were from south China, 335 were from east China, and 10 from Japan. In addition, 21 varieties with diverse morphological characteristics were used for first-round SSR marker screening (Table S2). The young leaves from 30 individual plants were collected for total genomic DNA extraction, using the cetyltrimethylammonium bromide (CTAB) method, as previously described (Tang et al., 2007).

2.2 SSR-PCR reaction

A total of 287 markers were selected from previous studies (Lowe et al., 2002; Wang et al., 2008; Ban, 2009; Cheng et al., 2009; Li, 2010; Liu et al., 2014; Song et al., 2015; Chen et al., 2017; Liu, 2017; Li et al., 2018; He et al., 2021) (Table S3). The selected SSR markers were labelled with 6-FAM (6-carboxyfluorescein), HEX (hexachlorofluorescein), ROX (6-carboxyl-X-rhodamine; passive reference dye), and TAMRA (5-carboxytetramethylrhodamine) fluorescent dyes at the 5′ end of the forward primer. The total volume of the polymerase chain reaction (PCR) was 20 μL, with a dNTP concentration of 0.20 mmol/L, and concentrations of forward and reverse primers of 0.25 μmol/L, 0.05 U/μL of *Taq* total genomic DNA polymerase, 1 × PCR buffer (containing Mg^{2+}, 2.5 mmol/l), and 50 ng/μL of DNA, and with the addition of

10.3389/fpls.2023.1112748

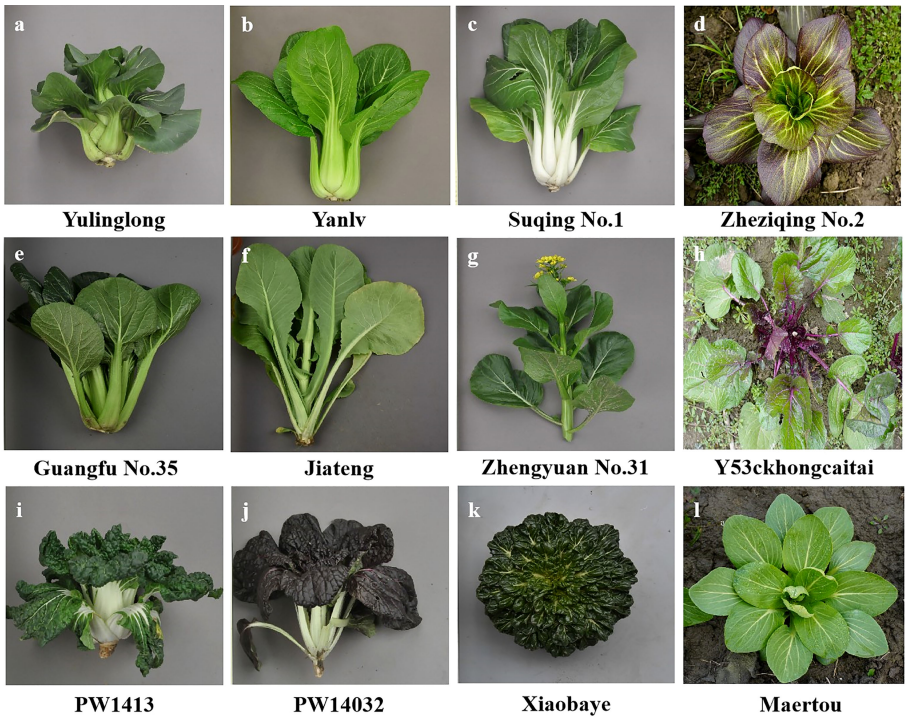


FIGURE 1
Different types of non-heading Chinese cabbage varieties: **(A–D)** pak-choi; **(E–H)** caitai; **(I–K)** tacai; and **(L)** duotoucai.

double-distilled H_2O up to a total of 20 μL. The PCR reaction conditions were as follows: pre-denaturation at 94°C for 5 minutes; denaturation at 94°C for 30 seconds, annealing at 55°C for 30 seconds, extension at 72°C for 45 seconds, for a total of 35 cycles; followed by extension at 72°C for 10 minutes; and then storage of the PCR reaction at 4°C.

2.3 Detection of PCR amplification products

During the first round of primer screening, primers were selected and detected by 6% polyacrylamide gel electrophoresis (PAGE) (Bao, 2015), with a constant power of 80 W; 2 μL of PCR product was added to each sample hole, and silver staining was performed after electrophoresis for 1–1.5 hours. Then the primers screened during the first round were labeled with different fluorescent dyes and were further screened and detected by a DNA analyzer (ABI3730).

2.4 Morphological evaluation

From December 2021 to March 2022, 423 non-heading Chinese cabbage varieties were planted at the Shanghai DUS testing base. In

TABLE 1 Details of morphological characteristics investigated in 423 non-heading Chinese cabbage varieties.

No	Characteristics	Type	Expression State (code)	H'	No	Characteristics	Type	Characteristic expression (code)	H'
1	Tiller	QL	absent(1); present(9)	0.04	16	Plant habit	QN	erect(1); erect to semi-erect(2);semi-erect(3); semi-erect to collapse(4); collapse(5)	0.98
2	Leaf hairiness	QL	absent(1)	0	17	Girdling	QN	absent(1); weak(2); medium(3); strong(4)	1.20
3	Leaf vein clarity	QL	weak(1); strong(2)	0.22	18	Plant height	QN	very low(1); very low to low(2); low(3); low to medium(4); medium(5); medium to high (6);high(7); high to very high(8); very high (9)	1.90
4	Inflorescence stem wax powders	QL	absent(1); present(9)	0.43	19	Plumpness of cabbage	QN	loose(1); medium(2); hard(3)	0.94
5	Seed coat color	PQ	yellow(1); brown(2); dark brown(3)	0.63	20	Leaf length	QN	very short(1); very short to short(2); short (3); short to medium(4); medium(5); medium to long(6); long(7); long to very long(8); very long(9)	1.90

(Continued)

TABLE 1 Continued

No	Characteristics	Type	Expression State (code)	H'	No	Characteristics	Type	Characteristic expression (code)	H'
6	Cotyledon color	PQ	light green(2); medium green (3); dark green(4); purple(5)	0.85	21	Leaf width	QN	very narrow(1); very narrow to narrow(2); narrow(3); narrow to medium(4); medium (5); medium to broad(6); broad(7); broad to very broad(8); very broad(9)	1.98
7	Leaf type	PQ	platy(1); divided leaf(2)	0.06	22	Leaf undulation of margin	QN	absent(1); 2(very weak)	1.39
8	Leaf shape	PQ	lanceolate (1); oval (2); elliptic(4); round oval(5); near round(6)	1.51	23	Leaf degree of blistering	QN	absent(1); very weak(2); weak(3); weak to medium(4); medium(5); medium to strong (6); strong(7); strong to very strong(8); very strong(9)	1.37
9	Leaf apex	PQ	blunt tip(1); circle(3); broad circle(4)	0.73	24	Leaf glossiness	QN	absent(1); weak(2); strong(3)	0.68
10	Leaf color	PQ	yellow green(1); light green (2); medium green(3); dark green(4); deep green(5); purple-red(6); purple(7)	1.34	25	Leaf number	QN	very less to less(2); less(3); less to medium (4); medium(5); medium to more(6); more (7); more to very more(8); very more(9)	1.77
11	Leaf margin features	PQ	inward(1); flat(2); outward (3)	0.42	26	Petiole thickness	QN	very thin(1); very thin to thin(2); thin(3); thin to medium(4); medium(5); medium to thick(6); thick(7); thick to very thick(8); very thick(9)	1.97
12	Petiole shape in horizontal section	PQ	subcircular(1); crescent(2); flat(3)	0.11	27	Petiole length	QN	very short to short(2); short(3); short to medium(4); medium(5); medium to long(6); long(7); long to very long(8); very long(9)	1.55
13	Petiole color	PQ	white(1); green white(2); light green(3); medium green (4); dark green(5); purple(6)	1.28	28	Petiole width	QN	very narrow(1); very narrow to narrow(2); narrow(3); narrow to medium(4); medium (5); medium to broad(6); broad(7); broad to very broad(8); very broad(9)	1.79
14	Inflorescence stem color	PQ	green(2); light green(3)	0.13	29	Bolting period	QN	very early to early(2); early(3); early to medium(4); medium(5); medium to late(6); late(7); late to very late(8); very late(9)	1.40
15	Flower color	PQ	white(1); light yellow(2); yellow(3); orange red(4)	0.38	30	Axillary bud generation ability	QN	absent or very weak(1); very weak to weak (2); weak(3); weak to medium(4); medium (5); medium to strong(6); strong(7); strong to very strong(8); very strong(9)	2.01

QL, qualitative characteristic; QN, quantitative characteristic; PQ, pseudo-qualitative characteristic; H′, Shannon–Wiener diversity index. Calculating by $H' = -\sum(Pi)\,(\ln Pi)$, where Pi is the proportion of individuals to total individuals of this species.

accordance with non-heading Chinese cabbage DUS test guidelines (http://www.nybkjfzzx.cn), 30 morphological characteristics were investigated (Table 1): four qualitative characteristics, 11 pseudo-qualitative characteristics, and 15 quantitative characteristics. The Shannon–Wiener diversity index of morphological characteristics was calculated as $H' = -\sum(Pi)\,(\ln Pi)$, where Pi is the proportion of individuals to total individuals of this species. The 'Pi' is an explanation of the formula, and the specific number of individuals depends on the expression state of the characteristics. were assigned a code from 1 to 9. For each characteristic of a variety, the expressed state was coded as 1 and the non-expressed state was coded as 0. The programming language R was used for 0 or 1 data format conversion, to build a 0/1 data matrix.

2.5 Data analysis

Raw electrophoresis data were read by SSR Analyzer V1.2.6 software (Wang et al., 2018). The genetic distance between different varieties was calculated by PowerMarker V3.25 software (Liu and Muse, 2005), and the unweighted pair group method with arithmetic means (UPGMA) clustering map based on Nei's genetic distance was constructed using MEGA5.0 software (Kumar et al., 2004). Genetic diversity parameters, including minor allele frequency (MAF), observed number of alleles (Na), observed heterozygosity (Ho), expected heterozygosity (He), polymorphic information content (PIC), and fixation index (Fst), were calculated using GenAlEx 6.51 software (Peakall and Smouse, 2012), which was also used for principal component analysis (PCA) and analysis of molecular variance (AMOVA). Combining morphological and molecular data, NTSYS2.11 software was used for genetic similarity analysis (James, 1987). Using qualitative data in the similarity module, the original 0/1 matrix generated by the morphological characteristic code and genotype data was adopted to calculate the genetic similarity (GS). The Mantel test was used to confirm the correlation between the similarity coefficient matrix generated from the morphological data and the SSR genotype data. Structure 2.34 software was used to analyze the population genetic structure from different regions of China (Falush et al., 2007). Assuming that the population number K was 1–10 and was

tested one by one, each *K*-value was estimated to be repeated 20 times: 5,000 iterations were not counted and the MCMC (Markov chain Monte Carlo) value was 50,000. The average value of ln*P* (*D*) was used for population estimation, the optimal population number was determined by the maximum likelihood method, and the corresponding *K*-value was calculated. Finally, we used NTSYS2.11 software to test the similarity of the phenotypic data of 14 varieties (five candidate varieties, with their corresponding similar varieties provided by applicants, and those screened by the SSR fingerprint database in this study).

3 Results

3.1 Establishment of variety identification system for non-heading Chinese cabbage

3.1.1 Core primer screening and polymorphism analysis

During the first round of primer screening, 21 representative varieties were used for PCR amplification, and 6% PAGE was used for electrophoresis (Figure 2A). As a result, 57 pairs of primers with high levels of polymorphism were screened. During the second round of primer screening, fluorescent dyes were labelled at the 5′ end of each of the 57 pairs of primers, and the fluorescent markers were used to amplify 96 varieties by capillary electrophoresis (Figures 2B–E). Based on the criteria of stable and simple fluorescence peak, low missing rate, high levels of polymorphism, and even distribution on chromosomes, 23 pairs of primers were finally selected as core primers, with the size of alleles ranging from 99 bp (SSR221) to 355 bp (SSR227). Detailed information on those primers is provided in Table 2.

Using 23 pairs of primers, 423 non-heading Chinese cabbage varieties were detected and a total of 314 alleles were obtained, with an average of 13.65 alleles per marker (Table 3). The variation range of MAF was 0.209 (SSR222) to 0.611 (SSR101), with an average of 0.419; the Ho amplitude ranged from 0.322 (SSR101) to 0.732 (SSR256), with an average of 0.530; the He amplitude was between 0.590 (SSR125) and 0.916 (SSR222); the PIC value ranged from 0.555 (SSR56) to 0.911 (SSR222), greater than 0.5, indicating high levels of polymorphism of all 23 markers; and the Fst of each molecular marker ranged from 0.045 (SSR221) to 0.547 (SSR266), with an average of 0.270. The above parameters showed that the 23 markers selected were high in polymorphism and could be used for genetic diversity detection among non-heading Chinese cabbage varieties, variety identification, and similar variety screening for the distinctness test.

3.1.2 Allelic sites calibration and database construction

According to the original capillary electrophoresis data, different allelic sites were named, and each allele's corresponding reference varieties were selected to calibrate systematic errors among different experimental batches or detection platforms. The size of allelic sites corresponding to each primer and the corresponding reference varieties are listed in Table S4.

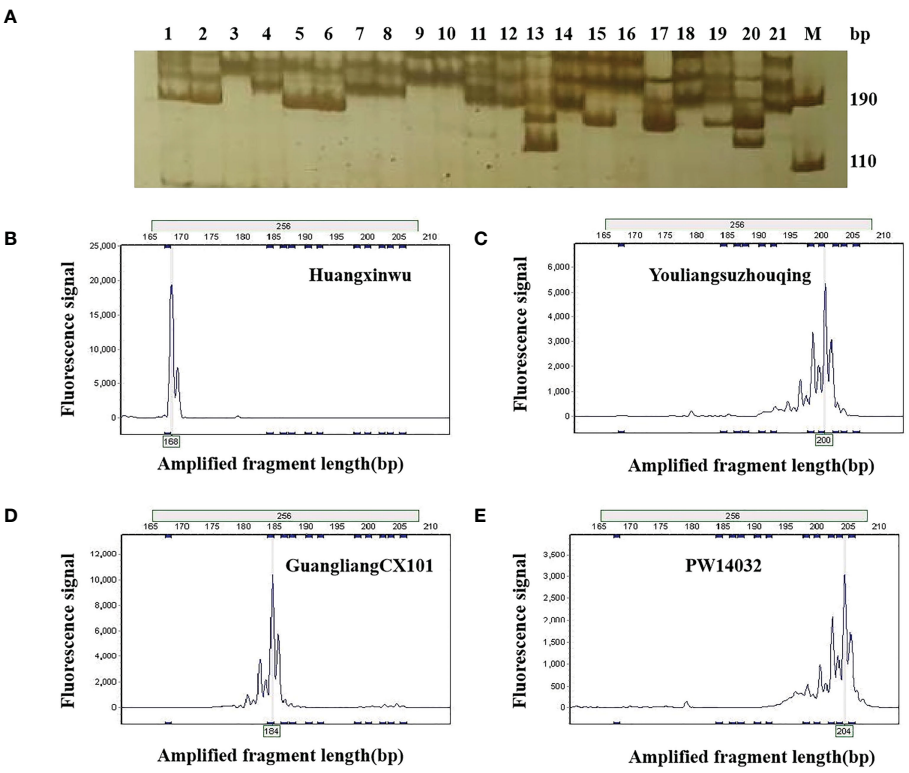


FIGURE 2
Allelic variation in 21 non-heading Chinese cabbage varieties using by primer SSR256. **(A)** Allelic variation in 21 varieties using PAGE. **M**, DNA marker; the number 1–21 in Figure 1a corresponds to the variety number in Table S2. **(B–E)** show the allelic variations in varieties 17, 9, 14, and 16 by fluorescence capillary electrophoresis. PAGE, polyacrylamide gel electrophoresis.

10.3389/fpls.2023.1112748

TABLE 2 Chromosome distribution and allelic variation range of the 23 primer sequences studied.

No.	Primer number	Sequence (5′→3′)	Chromosomal position	Allelic variation range (bp)
1	SSR101	F: TGGAGTGTTTGTTGTAAGCTCAA R: TTCGGGATGAGAGTTCCAAG	5	188–227
2	SSR125	F: TGCTCTTTGACACGTGCTATC R: AGAGGAGAGAAGGGGAGAGG	1	110–139
3	SSR136	F: TGATCACTGGGGTCCATTTA R: CTGCGTCGAAGTTAGAGACG	2	153–203
4	SSR138	F: TGCGTGCGGATTATCATCTA R: GGACGTAAACTTAGCACGATTC	2	160–174
5	SSR192	F: TAATCGCGATCTGGATTCAC R: ATCAGAACAGCGACGAGGTC	5	114–162
6	SSR198	F: GGTCAGGTGCTACTCAGACTCC R: TTGAAGAGGATCCACCAAAAG	3	276–314
7	SSR206	F: TGTCAGTGTGTCCACTTCGC R: AAGAGAAACCCAATAAAGTAGAACC	8	124–207
8	SSR207	F: TCAGCCTACCAACGAGTCATAA R: AAGGTCTCATACGATGGGAGTG	6	144–213
9	SSR22	F: ATGCACAGAGGAAGAAACCG R: GGGGATGAAGAAGAAGCAGA	1	155–191
10	SSR221	F: GTTCTCAAAGGGAAACCGAAAAACA R: GAGTTGGCCAGAGATTTACATGCGT	4	99–178
11	SSR222	F: CAAGAGCAAGTTTGAAACAAACGAT R: CATCAGTTCTTGATATGCTAGGTGA	6	175–280
12	SSR227	F: TTCCACCTCTCTGCTCCAAC R: ATGCGTGAGCGAGGATAACT	2	271–355
13	SSR228	F: GGAGTCCACTTCATGGAGGA R: CTCTTGCTCGTAGGTTTCCG	8	233–274
14	SSR229	F: TCAGTCACAAAAAGTCAACTCAAA R: ACGGAGTAGGAGTTGGGAGG	9	114–148
15	SSR238	F: TTTGACATCGTGCAATGCTA R: TTGGGCTGGTCCTGAAGATA	3	278–325
16	SSR247	F: GGTCCATTCCTTTTTGCATCTG R: CATGGCAAGGGGTAACAAACAT	7	128–154
17	SSR256	F: GGAGCCAGGAGAGAAGAAGG R: CCCAAAACTTCCAAGAAAAGC	3	168–206
18	SSR266	F: TCGGATTTGCATGTTCCTGA R: CCGATACACAACCAGCCAACT	7	187–305
19	SSR283	F: CCAACACCAAATCGCATAATC R: GGAGCTCCCACCTACAGTTTC	10	163–182
20	SSR45	F: GATTTGGGCCATTTGGATTA R: TTGAGCATTGTTCCCAGACA	4	206–230
21	SSR56	F: GTTAAGTTCGAACGCGAAGG R: GATCGGGGAAAATTAGGGAA	9	241–272
22	SSR66	F: ATTCAAAGACAAAGGAATGCCTGAG R: GTTTCTTTGATCCTGTCGAATGGCATTAATAAA	6	123–144
23	SSR90	F: TGCCTTTGTGTTCAGCTCAC R: CCCAAACGCTTTTGACACAT	10	202–211

bp, base pair; F, forward; R, reverse.

Based on the allelic sites data detected on the 423 non-heading Chinese cabbage varieties, the DNA molecular database was successfully constructed using SSR Analyzer V1.2.6 software. To improve the efficiency of database construction, 23 pairs of fluorescent primers were further divided into five groups (Table 4) according to the fluorescent color and amplified fragment size. Primers in each group could be mixed for multiple fluorescent capillary electrophoresis.

10.3389/fpls.2023.1112748

TABLE 3 Genetic parameters of the 23 SSR markers.

Marker	MAF	Na	Ho	He	PIC	Fst
SSR101	0.611	10	0.322	0.596	0.572	0.460
SSR125	0.610	14	0.386	0.590	0.561	0.346
SSR136	0.380	7	0.531	0.750	0.712	0.292
SSR138	0.335	8	0.723	0.794	0.767	0.089
SSR192	0.454	20	0.617	0.753	0.735	0.181
SSR198	0.599	12	0.490	0.616	0.598	0.204
SSR206	0.365	26	0.677	0.826	0.814	0.181
SSR207	0.527	6	0.336	0.618	0.557	0.455
SSR22	0.429	10	0.560	0.732	0.696	0.235
SSR221	0.490	16	0.693	0.725	0.708	0.045
SSR222	0.209	40	0.601	0.916	0.911	0.344
SSR227	0.233	23	0.461	0.850	0.834	0.458
SSR228	0.498	13	0.501	0.686	0.651	0.269
SSR229	0.474	9	0.546	0.654	0.596	0.164
SSR238	0.413	12	0.467	0.760	0.733	0.386
SSR247	0.346	7	0.579	0.701	0.644	0.174
SSR256	0.177	16	0.732	0.864	0.849	0.153
SSR266	0.415	30	0.356	0.787	0.771	0.547
SSR283	0.388	6	0.556	0.725	0.678	0.234
SSR45	0.456	7	0.414	0.671	0.614	0.383
SSR56	0.495	7	0.513	0.624	0.555	0.177
SSR66	0.283	10	0.551	0.780	0.744	0.293
SSR90	0.452	5	0.586	0.690	0.640	0.151
Mean	0.419	13.6	0.530	0.726	0.693	0.270

MAF, minor allele frequency; Na, observed number of alleles; Ho, observed heterozygosity; He, expected heterozygosity; PIC, polymorphic information content; Fst, fixation index.

TABLE 4 Grouping of 23 core primers according to different fluorescent-labelled colors.

Group Label	1	2	3	4	5
6-FAM	SSR125(110–139)	SSR138(150–174)	SSR256(164–206)	SSR221(99–178)	SSR136(153–203)
	SSR283(163–182)	SSR45(206–228)	SSR56(241–270)		SSR198(278–314)
	SSR228(233–274)	SSR227(271–355)			
ROX	SSR66(123–148)	SSR101(197–225)	SSR247(128–149)	SSR90(174–211)	SSR192(116–158)
	SSR266(188–305)	SSR238(278–325)			SSR222(175–227)
TAMRA		SSR229(114–148)	SSR22(155–194)	SSR206(124–207)	
HEX	SSR207(144–214)				

The selected SSR markers were labelled with 6-FAM, (6-carboxyfluorescein); HEX, (hexachlorofluorescein); ROX, (6-carboxyl-X-rhodamine; passive reference dye); or TAMRA, (5-carboxytetramethylrhodamine) fluorescent dyes at the 5′ end of the forward primer.

3.1.3 Discrimination power of core primers

To assess the accuracy and efficiency of core primers in distinguishing varieties, the 423 non-heading Chinese cabbage varieties were clustered based on the Nei's distance of 23 SSR markers (Figure 3). The clustering results showed that 418 out of the 423 varieties combinations could be distinguished by the 23 core primers, and that five groups could not be distinguished,because the genetic distance between varieties in each group was close to zero.

10.3389/fpls.2023.1112748

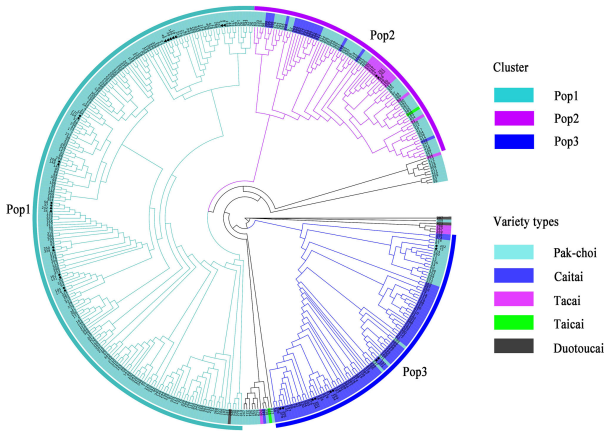


FIGURE 3
Cluster of 423 non-heading Chinese cabbage varieties based on Nei's distance of 23 simple sequence repeat (SSR) markers. ▲ represents two varieties with zero difference in the number of alleles; ★ represents two varieties with one difference in the number of alleles.

By using the formula [(423 * 422)/2 – 5]/(423 * 422)/2, the distinguishing rate for the 23 core SSR markers in 423 varieties was calculated to be 99.994%. By clustering, the 423 varieties could be divided into three main groups. In Pop1(n=226), besides duotoucais, the other varieties were pak-chois; Pop2(n=83) contained 52 pak-chois, 13 tacais, 17 caitais, and one taicai; and Pop3(n=93) comprised 19 pak-chois and 74 caitais. In addition, some local varieties were clustered separately, such as Shangwudong (412), Paopaoqing (286), and Xiangqingcai (308). Similarly, we conducted PCA to verify the clustering results, and principal coordinates 1 and 2 accounted for 19.39% and 7.13%, respectively, of the variation in the site information data (Figure 4). The AMOVA results showed that 66% of the variation came from within individuals. The genetic variation among individuals was greater than that among populations (Table S5). The fixation index (Fst) value was 0.134 ($p < 0.001$), indicating a high level of genetic differentiation among populations.

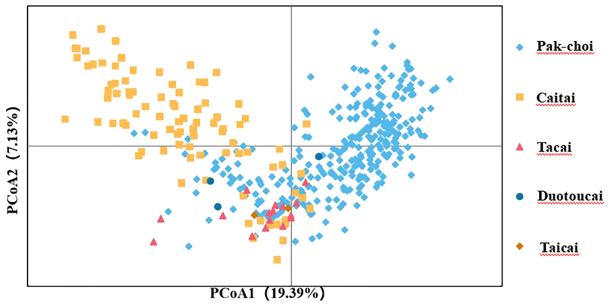


FIGURE 4
Principal component analysis (PCA) of 423 non-heading Chinese cabbage varieties based on 23 simple sequence repeat (SSR) markers. PCA results explained 19.39% and 7.13%, respectively, of the total variance by calculating the first two principal components.

3.2 Population structure analysis of the tested varieties

To explore the population distribution characteristics of non-heading Chinese cabbage, the population structure of 423 varieties was analyzed using the genotype data. The results showed that, for $K = 1–10$, the value of ln$P$ (D) increased with the increase in K-value (Figure 5A). A population structure distribution map based on Δk was constructed (Figure 5B), and the 423 varieties could be divided into three subgroups (Figure 5C). There were 84 varieties in subgroup I, from east China ($n = 70$), south China ($n = 6$), north China ($n = 1$), central China ($n = 4$), and Japan ($n = 3$); 138 varieties in subgroup II, from east China ($n = 113$), south China (n = 10), north China ($n = 7$), central China ($n = 5$), and Japan ($n = 3$); and 201 varieties in subgroup III, from east China ($n = 152$), south China ($n = 20$), north China ($n = 14$), central China ($n = 9$), northeast China ($n = 2$), and Japan ($n = 4$) (Figure 5C).

The population structure analysis showed that most genetic differences among non-heading Chinese cabbage varieties could be attributed to the geographical origins of the varieties. Varieties in subgroup I were mainly from east China, and varieties from east China also accounted for a large proportion of the other two subgroups; the caitai varieties were mainly from south China and clustered in subgroup II; and varieties from north China and northeast China were mainly clustered in subgroup III.

3.3 Correlation analysis between SSR markers and morphological characteristics

Descriptive statistics were based on 30 phenotypic characteristics of 423 non-heading Chinese cabbage varieties. The Shannon–Wiener diversity index of 30 characteristics ranged from 0 to 2.01, with an average of 1.03. In order to understand the relationship between SSR markers and morphological characteristics, the data from 30 morphological characteristics (markers) and those from the 23 core primers were converted into the 0/1 format, and the similarity coefficient of the two markers was calculated. The results showed that the similarity coefficient of the morphological markers and the SSR markers was moderate ($r = 0.53$) (Figure 6). Therefore, combining morphological and SSR markers would be more helpful for identifying non-heading Chinese cabbage.

Five groups not distinguished by the 23 core markers, 'Yanchun' and 'Yanlv', 'Guanmei No. 2' and 'Jinpin No. 3', 'Jingguan No. 1' and 'Huaxin', 'Jingguan No. 1' and 'Xinxiaqing No. 2', and 'Huaxin' and 'Xinxiaqing No. 2', were further compared through a field growing test. The plants in each group were very similar (Figure 7), although in each group slight differences were found in some visually observed characteristics such as seed coat color, plumpness of cabbage, leaf margin undulation, or bubble degree (Table S6). The variance analysis of six quantitative characteristics also revealed the existence of some differences in leaf length, leaf width, petiole length, and petiole thickness in four group varieties, but not in the group comprising 'Guanmei No. 2' and 'Jinpin No. 3' (Figure 7). These results indicated a certain degree of consistency in the identification of varieties between SSR markers and morphological characteristics. The

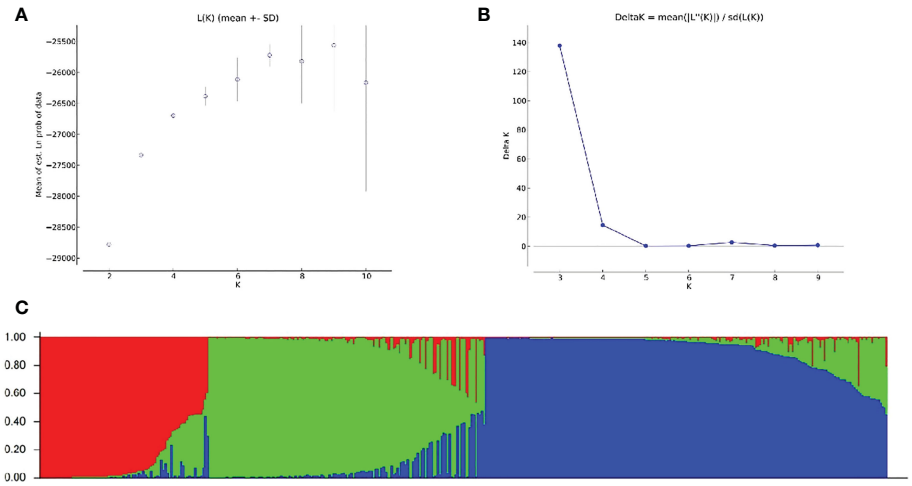


FIGURE 5
Population structure analysis of 423 non-heading Chinese cabbage varieties based on 23 simple sequence repeat (SSR) markers. **(A)** The mean value of ln*P* (D) was used to estimate the population structure, and the range of *K*-values was 1–10. **(B)** Using the curve of Δ*K* obtained by ln*P* (D), the optimal *K*-value was determined to be 3. **(C)** The 423 non-heading Chinese cabbage varieties studied clustered in three subgroups (subgroup I, red; subgroup II, green; and subgroup III, blue). Each histogram represents a variety in which different colors represent the estimated component coefficients using *Q*-values.

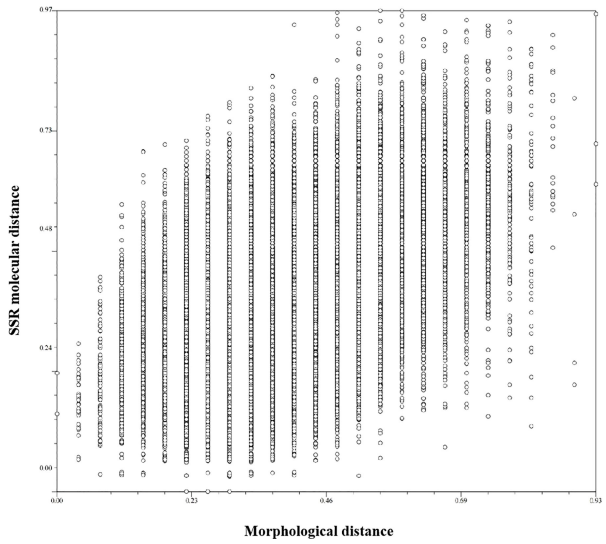


FIGURE 6
Comparison of morphological distance and molecular distance of 423 non-heading Chinese cabbage varieties. The abscissa is morphological distance, and the ordinate is molecular distance. The similarity coefficient is 0.53.

identification results based on morphological characteristics were more accurate and reliable than those based on SSR markers and, when used together with the molecular markers, could obviously improve identification efficiency.

3.4 Application of the SSR fingerprint database

In order to evaluate the application of the SSR fingerprint database in screening for similar varieties using the distinctness test, we selected five candidate varieties for which PVRs had been applied, and compared the similar varieties provided by the applicants (five varieties) with those screened by the SSR fingerprint database (five varieties) through the field planting test. (One of the varieties screened by the molecular fingerprint database was the same as the breeder provided, so there were 14 varieties.) The morphological characteristics comparison showed that four candidate varieties were similar to varieties screened by the SSR fingerprint database in this study (Figure 8). The morphological similarity between 'Huiwu No. 17' and 'Tadiwu No. 1' was 0.73, that between 'Huaerziqingfei' and 'Dongfangqinggeng' was 0.53, that between 'Heihuanghou' and 'Heimeigui' was 0.48, and that between 'Rehuo No. 16' and 'Jinpinxinxia' was 0.67.

As for candidate variety 'CT9970', it was more similar to the similar variety 'Biangubaicai' selected by the SSR fingerprint database than to the variety 'Lingxia 55' provided by its applicant. Breeding process analysis showed that 'CT9970' originated from 'Biangubaicai' and retained most of its morphological characteristics, whereas 'Lingxia 55' was the F_1 generation of 'CT9970' and 'CL45' (caitai variety), and resulted in low levels of similarity with 'CT9970'.

Thus, through morphological verification, the SSR fingerprint database can be used not only to screen similar varieties in the distinctness test, but also to preliminarily assess their genetic relationship.

4 Discussion

With the completion of whole-genome sequencing of non-heading Chinese cabbage, more SSR markers have been developed and utilized (Li et al., 2020). Because of the advantages of codominance and high levels of polymorphism, SSR markers provide an effective tool for studying the genetic diversity of non-heading Chinese cabbage (Li et al., 2021). In recent years, SSR markers have been widely used in predicting the genetic diversity

10.3389/fpls.2023.1112748

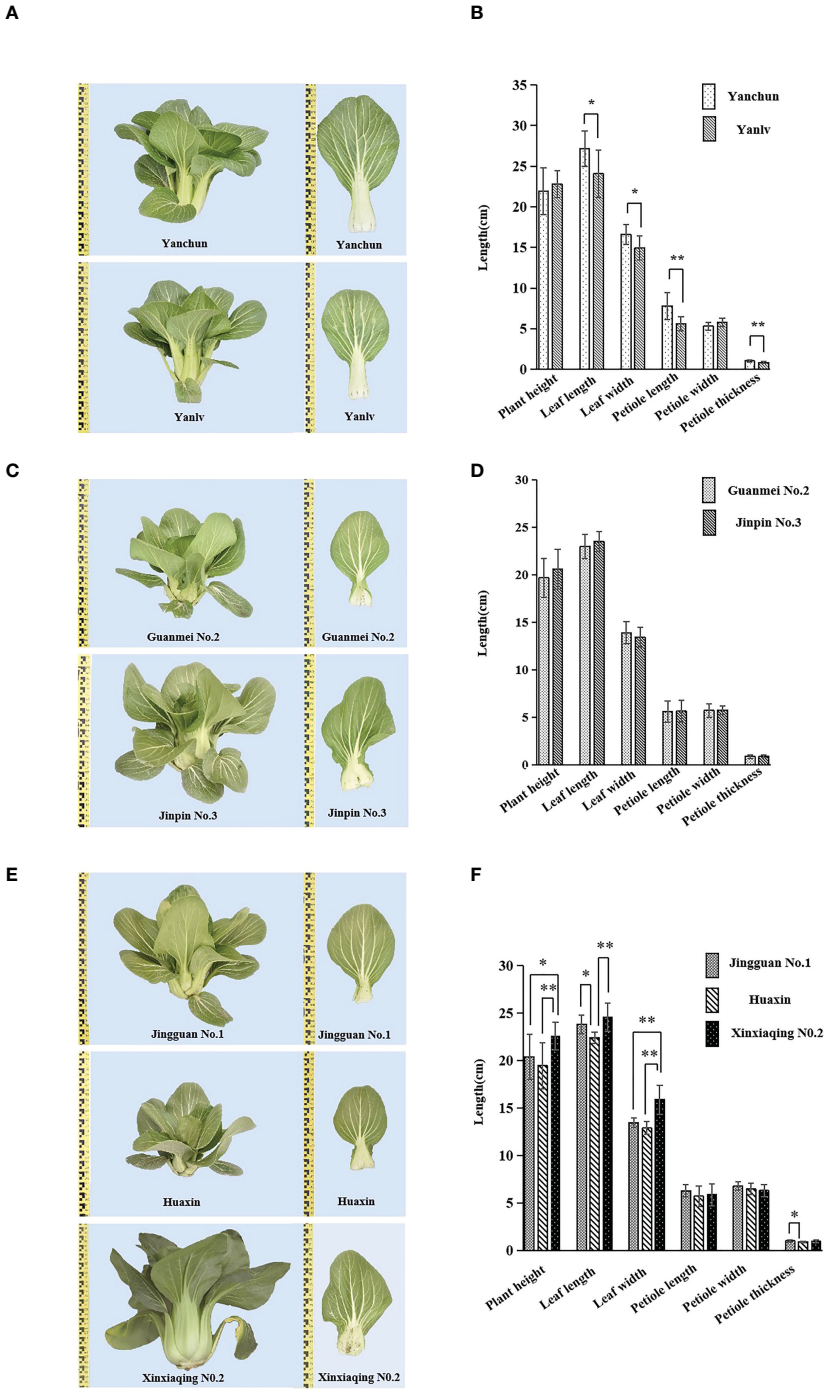


FIGURE 7
Phenotype comparison and quantitative characteristics ANOVA of five groups of varieties with no differences in simple sequence repeat (SSR) markers: 'Yanchun' and 'Yanlv' **(A, B)**; 'Guanmei No. 2' and 'Jinpin No. 3' **(C, D)**; 'Jingguan No. 1' and 'Huaxin', 'Jingguan No. 1' and 'Xinxiaqing No. 2', and 'Huaxin' and 'Xinxiaqing No. 2' **(E, F)**. Ten plants were used for the analysis (*significant at $p < 0.05$; **highly significant at $p < 0.01$).

and germplasm identification of non-heading Chinese cabbage germplasm resources (Han et al., 2008; Ma, 2015; Li et al., 2017; Li et al., 2018). In this study, by using 423 non-heading Chinese cabbage varieties with rich and diverse phenotypes, 23 pairs of SSR primers (out of 287 analyzed) with better performance than those used in previous studies (Xue et al., 2014; Yang et al., 2020) were identified (average PIC value of 0.693). This may be attributed to the large number of varieties collected, their rich genetic diversity, and the highly accurate capillary electrophoresis detection method used in this study.

Clustering results in this study showed that most of the 423 non-heading Chinese cabbage varieties fell into one of three main groups, pak-choi, caitai, and tacai, which was in line with the actual status of breeding and production. In the study of Ma et al. (2015), Pak-choi

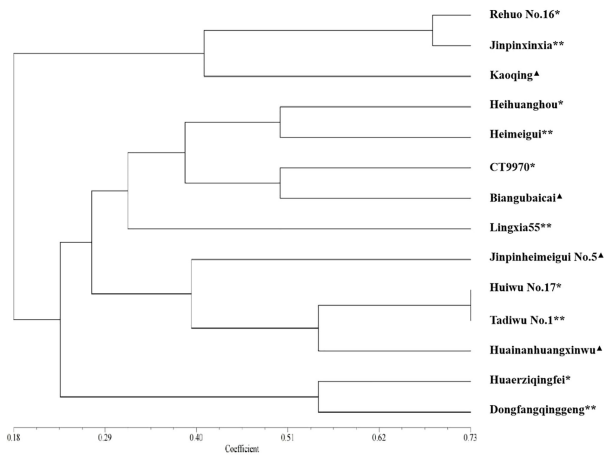


FIGURE 8
Similarity analysis of candidate varieties and their corresponding similar varieties based on 30 morphological characteristics. The scale at the bottom is the similarity coefficient. *: candidate varieties; **: varieties screened by SSR fingerprint database; ▲: the applicant submitted similar varieties.

varieties are clustered with Caitai, and Tacai in different degrees, which was similar to the research results of this study. In addition, PCA in this study also showed that there was obvious interspecific crossing and extensive gene exchange between the pak-choi and caitai genetic backgrounds (Figure 4), but this phenomenon has been seldom mentioned in previous studies.

According to a previous study, purple is not completely dominant over green in the inheritance of non-heading Chinese cabbage, and the purple color largely depends on anthocyanin content (Zhu, 2017). However, we observed that the hybrid progeny of crosses between a green and a purple non-heading Chinese cabbage variety showed a distribution that was largely skewed towards the phenotype of purple parent, suggesting that all varieties of purple non-heading Chinese cabbage are likely to have the same genetic background.

The genetic diversity of non-heading Chinese cabbage was related to geographical origin (Wang et al., 2008; Liu et al., 2014). Population structure analysis in this study showed that, in east China, germplasm resources were more abundant and genetic diversity was greater, and the three provinces of Jiangsu, Zhejiang, and Fujian in east China had relatively independent genetic structures, which confirmed that non-heading Chinese cabbage in China might originate from the Jianghuai area (Cao et al., 1997).

Recent studies have shown that different varieties can be effectively distinguished and analyzed through complementary differences in morphological markers and molecular markers (Lee and Park, 2017). This complementary method is usually used in germplasm identification (Delfini et al., 2007; Haliloglu et al., 2022) and genetic diversity analysis (Guo et al., 2020; Chikh-Rouhou et al., 2021). Theoretically, one morphological characteristic would be usually regulated by multiple genes. In this study, only five groups could not be distinguished by the 23 core markers, and the field growing comparison test showed that varieties in each of the five groups were very similar but were still distinguishable by some visually observed or measured characteristic. Molecular markers correlated to a medium extent (r = 0.53) with morphological characteristics, which was higher than that in a previous study on peanuts (0.347) (Hong et al., 2021), but no functional molecular markers associated with morphological characteristics in non-heading Chinese cabbage were found in this study. Therefore, without enough functional markers, molecular markers cannot completely replace morphological markers, but a combination of both types of markers would be more accurate and efficient in variety identification and in similar variety screening for the distinctness test.

5 Conclusion

In this study, 23 out of 287 SSR markers were selected as the core markers, with an average PIC value of 0.693 and an average number of alleles of 13.65. Based on the 23 core markers, the SSR fingerprint database comprising 423 non-heading Chinese cabbage varieties was constructed, in which 418 out of the 423 varieties could be distinguished with a discrimination rate of 99.994%. The SSR fingerprint database constructed in this study could be used not only in the identification of varieties but also for similar varieties screening of distinctness test.

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

ZX and RH designed and supervised this project. JY and HZ performed most of the experiments. XW, YM, and JZ carried out part of the experiments. JY, HZ, and XW participated in data analysis. JY, HZ, XW, RH, and ZX wrote the manuscript. YL, GS, and HC revised the manuscript. RH and ZX supervised the study and revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The study was co-financed by the National Species Resources Protection Project (h20210472) and the Construction of Agricultural Products Quality and Safety Standards System (2130109).

Conflict of interest

Author GS is employed by Fujian Jinpin Agricultural Technology Co., Ltd.

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.2023.1112748/full#supplementary-material

References

Ban, Q. Y. (2009). *Construction of a genetic linkage map and mapping QTL of bolting character using F2 population in non-heading Chinese cabbage* (China (Nanjing: Nanjing Agricultural University).

Bao, S. Y. (2015). Improvement of DNA denaturing polyacrylamide gel electrophoresis in experiment teaching. *Exp. Sci. Technol.* 13, 122–124. doi: 10.3969/j.issn.1672-4550.2015.02.040

Cao, J. S., Cao, S. C., Miu, Y., and Lu, G. (1997). Cladistic operational analysis and study on the evolution of Chinese cabbage groups (*Brassica campestris l.*). *Acta Hortic. Sinica* 24, 35–42. doi: 10.3321/j.issn:0513-353X.1997.01.007

Cheng, Y., Geng, J. F., Zhang, J. Y., Wang, Q., Ban, Q. Y., and Hou, X. L. (2009). The construction of a genetic linkage map of non-heading Chinese cabbage (Brassica campestris ssp. *Chinensis Makino) J. Genet. Genomics* 36, 501–508. doi: 10.1016/s1673-8527(08)60140-x

Chen, J. F., Li, R. H., Xia, Y. S., Bai, G. H., Guo, P. G., Wang, Z. L., 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, e0184736. doi: 10.1371/journal.pone.0184736

Chikh-Rouhou, H., Mezghani, N., Mnasri, S., Mezghani, N., and Garcés-Claver, A. (2021). Assessing the genetic diversity and population structure of a tunisian melon (Cucumis melo l.) collection using phenotypic traits and SSR molecular markers. *Agronomy* 11, 1121. doi: 10.3390/agronomy11061121

Chu, Y. X., Deng, S., Li, S. G., Liu, D., Chen, H. R., Ren, L., et al. (2020). Screening and application of SSR molecular markers in varieties identification of flower vegetables. *Mol. Plant Breeding* 20, 163–175. doi: 10.13271/j.mpb.020.000163

Delfini, J., Moda-Cirino, V., Ruas, C. F., Neto, J. D. S., Ruas, P. M., Buratto, J. S., et al. (2007). Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7, 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Guo, X., Cheng, F., and Zhong, Y. (2020). Genetic diversity of paeonia rockii (flare tree peony) germplasm accessions revealed by phenotypic traits, EST-SSR markers and chloroplast DNA sequences. *Forests* 11, 672. doi: 10.3390/f11060672

Haliloglu, K., Turkoglu, A., Tan, M., and Poczai, P. (2022). SSR-based molecular identification and population structure analysis for forage pea (Pisum sativum var. arvense l.) landraces. *Genes* 13, 1086. doi: 10.3390/genes13061086

Han, J. M., Hou, X. L., Xu, H. M., Shi, G. J., and Wang, J. J. (2008). RAPD analysis of genetic diversity of non-heading Chinese cabbage (Brassica campestris ssp.chinensis makino)germplasm. *J. Nanjing Agric. University* 31, 31–36. doi: 10.7685/j.issn.1000-2030.2008.03.006

He, X. L., Yang, D. Q., Du, Z. J., Shang, C. Y., and Zhong, F. L. (2021). Association analysis of morphological traits and SSR genetic diversity in non-heading Chinese cabbage. *Mol. Plant Breeding* 19, 1919–1927. doi: 10.13271/j.mpb.019.001919

Hong, Y. B., Pandey, M. K., Lu, Q., Liu, H., Gangurde, S. S., Li, S. X., et al. (2021). Genetic diversity and distinctness based on morphological and SSR markers in peanut. *Agron. J.* 6, 113. doi: 10.1002/agj2.20671

Hou, X. L., Li, Y., and Huang, Y. F. (2020). New advances in molecular biology of main characters and breeding technology in non-heading Chinese cabbage(Brassica campestris ssp. chinensis). *Acta Hortic. Sinica* 47, 1663–1677. doi: 10.16420/j.issn.0513-353x.2020-0534

Hou, X. L., and Song, X. M. (2012). Research and utilization of brassica campestris ssp. chinensis makino(non-heading Chinese cabbage) germplasm resources. *J. Nanjing Agric. University* 35, 35–42. doi: 10.7685/j.issn.1000-2030.2012.05.005

James, R. F. (1987). NTSYS-pc: microcomputer programs for numerical taxonomy and multivariate analysis. *Am. Statistician* 41, 330. doi: 10.2307/2684761

Kumar, S., Tamura, K., and Nei, M. (2004). MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings Bioinf.* 5, 150–163. doi: 10.1093/bib/5.2.150

Lee, O. N., and Park, H. Y. (2017). Assessment of genetic diversity in cultivated radishes (Raphanus sativus) by agronomic traits and SSR markers. *Scientia Horticulturae* 223, 19–30. doi: 10.1016/j.scienta.2017.05.025

Li, X. (2010). *Analysis of population structure and construction of genetic linkage map in non-heading Chinese cabbage* (China (Hebei: Agricultural University of Hebei).

Li, G. H., Chen, H. C., Zhan, Y., and Li, T. Y. (2017). Genetic diversity and phylogenetic relationships analysis of Chinese cabbage germplasm resources by SRAP and SSR. *Guangdong Agric. Sci.* 44, 37–45. doi: 10.16768/j.issn.1004-874X.2017.05.007

Li, Y., Liu, G. F., Ma, L. M., Liu, T. K., Zhang, C. W., Xiao, D., et al. (2020). A chromosome-level reference genome of non-heading Chinese cabbage [Brassica campestris (syn. brassica rapa) ssp. chinensis. *Hortic. Res.* 7 (1), 212. doi: 10.1038/s41438-020-00449-z

Li, P. R., Su, T. B., Zhao, Y. Y., Wang, W. H., Zhang, D. S., Yu, Y. J., et al. (2021). Assembly of the non-heading pak choi genome and comparison with the genomes of heading Chinese cabbage and the oilseed yellow sarson. *Plant Biotechnol. J.* 19, 966–976. doi: 10.1111/pbi.13522

Liu, K. Y. (2017). *Construction of genetic linkage map of non-heading Chinese cabbage and identification of plant type phenotypic traits* (China (NanJing: Nanjing agricultural university).

Liu, L. J., Liu, Z. C., Chen, H. R., and Zhou, L. G. (2012). SRAP markers and morphological traits could be used in test of distinctiveness, uniformity, and stability (DUS) of lettuce (Lactuca sativa) varieties. *J. Agric. Sci.* 4, 227. doi: 10.5539/jas.v4n3p227

Liu, K., and Muse, S. V. (2005). PowerMarker: An integrated analysis environment for genetic marker analysis. *Bioinformatics* 21, 2128–2129. doi: 10.1093/bioinformatics/bti282

Liu, T. K.Nanjing agricultural university (2021). *A molecular marker method for identification of non-heading chinese cabbage suzhouqing, aijiaohuang and wutacai*, China patent CN 108165647B. Nanjing, State Intellectual Property Office of the People 's Republic of China.

Liu, D. Y., Wang, X. H., Liu, Y., Zhang, J., and Chen, H. Y. (2014). Analysis of genetic diversity and relationship of pakchoi accessions based on SSR markers. *Mol. Plant Breeding* 12, 499–508. doi: 10.13271/j.mpb.012.000499

Li, G. G., Zhang, H., Zheng, Y. S., and Li, R. H. (2018). The genetic diversity analysis of Chinese flowering cabbage resources based on SSR marker. *Genomics Appl. Biol.* 37, 1257–1264. doi: 10.13417/j.gab.037.001257

Lowe, A. J., Jones, A. E., Raybould, A. F., Trick, M., Moule, C. L., and Edwards, K. J. (2002). Transferability and genome specificity of a new set of microsatellite primers among brassica species of the U triangle. *Mol. Ecol. Notes* 2, 7–11. doi: 10.1046/j.1471-8286.2002.00126.x

Ma, J. J. (2015). *Genetic diversity analysis of non-heading chinese cabbage using SSR markers and agronomic traits* (China (Yangling: Northwest A&F University).

Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: Genetic analysis in excel. population genetic software for teaching and research–an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460

Song, X. M., Ge, T. T., Li, Y., and Hou, X. L. (2015). Genome-wide identification of SSR and SNP markers from the non-heading Chinese cabbage for comparative genomic analyses. *BMC Genomics* 16, 328. doi: 10.1186/s12864-015-1534-0

Tang, W. K., Tan, X., Zhang, H., Huang, G. Q., Xu, W. L., and Li, X. B. (2007). A rapid and simple method of DNA extraction from plant samples. *J. Huazhong Normal University* 3, 441–449. doi: 10.3321/j.issn:1000-1190.2007.03.030

Wang, F. G., Li, X., Yang, Y., Yi, H. M., Jiang, B., Zhang, X. C., et al. (2018). SSR analyser: A special software suitable for SSR fingerprinting of plant varieties. *Scientia Agricultura Sinica* 51, 2248–2262. doi: 10.3864/j.issn.0578-1752.2018.12.003

Wang, X. Y., Yu, S. C., Zhang, F. L., Yu, Y. J., Zhao, Y. Y., and Zhang, D. S. (2008). SSR fingerprinting and genetic distinctness of pak-choi (Brassica rapa l.ssp.chinensis Makino).ACAT. *Agric. Boreali-Sinica* 5, 97–103. doi: 10.7668/hbnxb.2008.05.021

Xue, S. X., Li, X., Gu, A. X., Zhao, J. J., Wang, Y. H., and Shen, S. X. (2014). Cluster analysis of non-heading chinese cabbage cultivars based on SSR marker. *Northern Horicult.* 3, 83–87.

Yang, D. Q., He, X. L., Du, Z. J., Wang, S. B., Sun, L. W., Zhang, L. Y., et al. (2020). Development and polymorphism analysis of EST-SSR markers based on transcriptome of non-heading Chinese cabbage (*Brassica rapa ssp.chinensis*). *J. Agric. Biotechnol.* 28, 13–21. doi: 10.3969/j.issn.1674-7968.2020.01.002

Yan, X. L., Guan, Z. R., Wen, W., Zhang, Z. F., Wang, C. Y., Shen, J. J., et al. (2021). Establishment and application of mustard variety identification system based on SSR markers(*Brassica juncea l.*). *J. Plant Genet. Resourc.* 22, 758–770. doi: 10.13430/j.cnki.jpgr.20201014002

Yu, S. C.Beijing Academy of Agriculture and Forestry Sciences (2014). *A set of SSR primer combinations suitable for constructing nucleic acid fingerprinting database of non-heading chinese cabbage and its application*, China patent CN 104073561B. Beijing, State Intellectual Property Office of the People 's Republic of China.

Zhan, H., Wang, D. J., Sun, J. M., Zheng, Y. S., Yao, F. X., Xu, J. F., et al. (2014). Development and application of a high-throughput Chinese cabbage DNA profiling system based on SSR markers. *J. Plant Genet. Resourc.* 15, 815–823. doi: 10.13430/j.cnki.jpgr.2014.04.020

Zhou, Y., Cao, H. H., Wang, Y., Huang, X. C., Gu, S. G., and Fu, L. J. (2020). Application prospect of molecular markers in DUS testing. *Anhui Agric. Sci. Bull.* 26, 15–18. doi: 10.3969/j.issn.1007-7731.2020.17.006

Zhu, H. F. (2017). *Study of purple inheritance, response to light and temperature and new germplasm creatation in brassica campestris.ssp.chinensis makino* (China (Shanghai: Shanghai Jiao Tong Universit).

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 20 March 2023
DOI 10.3389/fpls.2023.1116851

OPEN ACCESS

EDITED BY
Satoshi Watanabe,
Saga University, Japan

REVIEWED BY
Fabio Palumbo,
University of Padua, Italy
Ruslan Kalendar,
University of Helsinki, Finland

*CORRESPONDENCE
Christine D. Chase
cdchase@ufl.edu

†PRESENT ADDRESSES
Melinda R. Grosser,
Department of Biology, University of North Carolina, Asheville, NC, United States
Samantha K. Sites,
North Carolina Department of Health and Human Services, Raleigh, NC, United States
Mayara M. Murata,
Stricto Sensu Department, Universidade Norte do Paraná, Londrina, Paraná, Brazil
Kyra Love Harriage,
Career and Technical Education Department, Polk County Public Schools, Bartow, FL, United States

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

RECEIVED 05 December 2022
ACCEPTED 02 March 2023
PUBLISHED 20 March 2023

CITATION
Grosser MR, Sites SK, Murata MM, Lopez Y, Chamusco KC, Love Harriage K, Grosser JW, Graham JH, Gmitter FG Jr. and Chase CD (2023) Plant mitochondrial introns as genetic markers - conservation and variation.
Front. Plant Sci. 14:1116851.
doi: 10.3389/fpls.2023.1116851

Plant mitochondrial introns as genetic markers - conservation and variation

Melinda R. Grosser[1†], Samantha K. Sites[1†], Mayara M. Murata[2†], Yolanda Lopez[3], Karen C. Chamusco[1], Kyra Love Harriage[1†], Jude W. Grosser[2], James H. Graham[2], Fred G. Gmitter Jr.[2] and Christine D. Chase[1*]

[1]Horticultural Sciences Department, University of Florida, Gainesville, FL, United States, [2]Citrus Research and Education Center, University of Florida, Lake Alfred, FL, United States, [3]Agronomy Department, University of Florida, Gainesville, FL, United States

Plant genomes are comprised of nuclear, plastid and mitochondrial components characterized by different patterns of inheritance and evolution. Genetic markers from the three genomes provide complementary tools for investigations of inheritance, genetic relationships and phenotypic contributions. Plant mitochondrial genomes are challenging for universal marker development because they are highly variable in terms of size, gene order and intergenic sequences and highly conserved with respect to protein-coding sequences. PCR amplification of introns with primers that anneal to conserved, flanking exons is effective for the development of polymorphic nuclear genome markers. The potential for plant mitochondrial intron polymorphisms to distinguish between congeneric species or intraspecific varieties has not been systematically investigated and is possibly constrained by requirements for intron secondary structure and interactions with co-evolved organelle intron splicing factors. To explore the potential for broadly applicable plant mitochondrial intron markers, PCR primer sets based upon conserved sequences flanking 11 introns common to seven angiosperm species were tested across a range of plant orders. PCR-amplified introns were screened for indel polymorphisms among a group of cross-compatible *Citrus* species and relatives; two *Raphanus sativus* mitotypes; representatives of the two *Phaseolus vulgaris* gene pools; and congeneric pairs of *Cynodon*, *Cenchrus*, *Solanum*, and *Vaccinium* species. All introns were successfully amplified from each plant entry. Length polymorphisms distinguishable by gel electrophoresis were common among genera but infrequent within genera. Sequencing of three introns amplified from 16 entries identified additional short indel polymorphisms and nucleotide substitutions that separated *Citrus*, *Cynodon*, *Cenchrus* and *Vaccinium* congeners, but failed to distinguish *Solanum* congeners or representatives of the *Phaseolus vulgaris* major gene pools. The ability of primer sets to amplify a wider range of plant species' introns and the presence of intron polymorphisms that distinguish congeners was confirmed by in silico analysis. While mitochondrial intron variation is limited in comparison to nuclear introns, these exon-based primer sets provide robust tools for the amplification of

mitochondrial introns across a wide range of plant species wherein useful polymorphisms can be identified.

KEYWORDS

group II intron, indel polymorphism, organelle genome, PCR-based markers, plant mitochondria, single nucleotide polymorphism

Introduction

Plant genetic information is distributed among nuclear, plastid and mitochondrial genomes (Mahapatra et al., 2021; Camus et al., 2022), and genetic markers for each genome provide complementary tools for investigations of inheritance and evolution (Qiu et al., 1999; Duminil and Besnard, 2021; Besse, 2021; Camus et al., 2022). Although genome sequencing is the gold standard for such studies, convenient PCR-based markers retain utility and appeal (Egan et al., 2012; Hodel et al., 2016; Besse, 2021). The distinctive features of each genome necessitate different strategies for marker development and create different opportunities for application. Plant plastid genomes are relatively conserved in gene order, moderately conserved in coding sequences and more polymorphic with respect to introns and intergenic spacers (Wicke et al., 2011), facilitating the development of universal primers for the PCR amplification and subsequent characterization of more variable regions. Amplified plastid sequences such as *rbc*L, *mat*K are the core of the DNA barcoding approach for distinguishing plant species (CBOL Plant Working Group, 2009), with the BOLD database facilitating applications (Ratnasingham and Hebert, 2007), and with intergenic spacer regions proving more variable and useful in distinguishing closer relatives (Shaw et al., 2014). Barcoding is also being combined with plastid genome sequencing for broader applicability and enhanced resolution (Tonti-Filippini et al., 2017). Plant mitochondrial markers provide an important adjunct to plastid DNA markers (Duminil and Besnard, 2021). Mitochondrial genotype can have a significant influence on plant phenotype (Bock et al., 2014; Colombatti et al., 2014; Hu et al., 2014; Dourmap et al., 2020). It is therefore important to be able to track mitochondrial contributions in sexual crosses and somatic cell fusions. Both plastid and plant mitochondrial genomes have uni-parental inheritance patterns, but these are not always concordant with respect to parent of origin (Camus et al., 2022). Moreover, horizontal gene transfer, observed in both organelle genomes, is especially prevalent in plant mitochondria (Keeling, 2009; Archibald and Richards, 2010). Plant mitochondrial gene coding sequences are, with some exceptions (Mower et al., 2007), highly conserved (Wolfe et al., 1987; Drouin et al., 2008), but genome size, gene order and intergenic sequences vary exensively between, and even within, plant species (Sloan, 2013; Gualberto and Newton, 2017). Mitochondrial restriction fragment length polymorphisms (RFLPs) are therefore readily detected within plant species (Levings and Pring, 1977; Palmer and Herbon, 1988), but the development of polymorphic PCR-based mitochondrial markers that work across a wide range of plant species is problematic. The lack of conserved gene order precludes the development of universal primer sets that will anneal to conserved coding sequences and amplify the highly polymorphic intergenic sequences.

Minisatellites and microsatellites (tandem repeats of 10 to 100, or less than 10 base pairs, respectively) identified within the sequenced mitochondrial genomes of some plant species have provided the basis for PCR-based polymorphic markers. Minisatellite repeat number polymorphisms have demonstrated intraspecific variation in *Beta vulgaris*, *B. maritima* (Nishizawa et al., 2000; Nishizawa et al., 2007), *Picea abies* (Sperisen et al., 2001; Bastien et al., 2003), *Pinus banksiana* (Godbout et al., 2005), and *Pinus ponderosa* (Mitton et al., 2000), as well as interspecific polymorphisms in *Brassica* and *Oryza* species (Honma et al., 2011). Interspecific, but not intraspecific, variation for a G_n microsatellite is present in the genus *Pinus* (Soranzo et al., 1999), whereas a compound, highly polymorphic microsatellite region reveals both intra- and interspecific variation in *Abies* (Jaramillo-Correa et al., 2013). Tandemly repeat mitochondrial loci are not generally conserved across diverse plant taxa and are not always polymorphic between related taxa, but recent work has identified extensive mitochondrial microsatellites among plant species (de Freitas et al., 2022; Xiong et al., 2022). These studies and databases of plant mitochondrial microsatellite repeats (Kumar et al., 2014; Sablok et al., 2015) facilitate the experimental search for loci that are polymorphic in specific taxa.

Plant mitochondrial introns present an under-explored approach for the development of more universal, PCR-based mitochondrial genome markers. PCR amplification of polymorphic introns with primers designed to conserved flanking exon sequences (Lessa, 1992) has allowed the development of nuclear genome markers in plant species having limited genomic information (Gupta et al., 2011; Li et al., 2012; Chandra et al., 2013; Kim et al., 2015) or limited genetic variability (Wang et al., 2010; Galeano et al., 2012). Angiosperm mitochondrial genomes encode 20-24 group II introns. Although sporadic intron loss is observed among evolutionary lineages, many of these introns are common to the sequenced angiosperm mitochondrial genomes (Kubo and Mikami, 2007), and flanked by conserved coding sequences that can be exploited for universal primer development. Laroche et al. (1997) surveyed the genomic sequences of six mitochondrial introns that were located within five genes and were common to five different angiosperm species and concluded that plant mitochondrial introns could provide a source of polymorphic

markers. Across these species, base substitutions per site were higher within introns than within exons. Insertion-deletion (indel) polymorphisms were observed at 0.2-0.5 times the frequency of base substitutions. These sequence comparisons were made across a small set of diverse angiosperm genera, and so did not determine whether plant mitochondrial introns are commonly polymorphic between cross-compatible species or within species – situations in which polymorphisms could function as useful genetic markers. These points require investigation as correct splicing of plant organelle group II introns depends upon a complex intron secondary structure and upon RNA-protein interactions with multiple, co-evolving, nuclear-encoded splicing factors (Bonen, 2008; de Longevialle et al., 2010; Brown et al., 2014) – requirements that potentially constrain the degree of intron polymorphism that can be found among close relatives.

DNA markers based upon PCR-amplified plant mitochondrial intron sequences have proved useful in some cases. While most plant mitochondrial microsatellite and minisatellite repeats are located in intergenic regions, polymorphic examples are found within introns (Sperisen et al., 2001; Godbout et al., 2005; Jaramillo-Correa et al., 2013; Potter et al., 2013; Xiong et al., 2022). Duminil et al. (2002) designed primer pairs for the amplification of 16 different introns, based upon the mitochondrial genome sequences of *Arabidopsis thaliana* and *Beta vulgaris*. These primer sets amplify their corresponding introns in 20-28 of 28 diverse angiosperm species, and some have been investigated for polymorphisms in related species. The PCR amplified NADH dehydrogenase subunit 1 intron 2 (*nad1*i2), NADH dehydrogenase subunit 4 intron 1 (*nad4*i1) and intron 2 (*nad4*i2) are not polymorphic within *Quercus robur*, but distinguish between *Q. robur* and *Q. rubra* (Demesure et al., 1995). Notably, complex mitochondrial SSR loci analyzed across 88 genomes are especially prevalent in the introns of *nad2*, *nad4* and *nad7* genes (Xiong et al., 2022).

Mitochondrial intron polymorphisms also have utility in citrus breeding and genetics. Commercial citrus types are complex hybrids with at least three maternal lineages among them - *Citrus maxima* (pummelo), *C. reticulata* (mandarin) and *C. medica* (citron). The genus overall has complex taxonomy (Moore, 2001; Wu et al., 2018). Froelicher et al. (2011) amplified short, internal, regions of *Citrus* NADH dehydrogenase subunit 2 intron 3 (*nad2*i3), NADH dehydrogenase subunit 5 intron 2 (*nad5*i2), and NADH dehydrogenase subunit 7 intron 1 (*nad7*i1), with primers based upon *A. thaliana* and *B. vulgaris* mitochondrial genome sequences. *Citrus* and citrus relatives are polymorphic for indels in these introns, which collectively identify seven *Citrus* mitotypes. Intron-flanking primers designed from alignment of conserved DNA sequences flanking introns common to seven sequenced angiosperm mitochondrial genomes (Grosser, 2011) generate intron amplification products that distinguish *C. maxima* from *C. reticulata* (Satpute et al., 2015) and *C. maxima* from *C. japonica* (kumquat) (Omar et al., 2017). Here, we demonstrate the utility of these primer sets for amplification of their target introns not only in the previously studied *C. maxima*, *C. reticulata* and *C. Japonica* lineages, but also across diverse angiosperm species. We further investigate the amplified introns for indel and single nucleotide polymorphisms (SNPs) that distinguish mitochondrial genomes within a plant species or between congeneric plant relatives, wherein polymorphic mitochondrial markers have potential applications in studies of evolution and inheritance.

Materials and methods

Plant materials and DNA extraction

The plant materials used in this study (Table 1) were selected to explore primer amplification across across six angiosperm orders and to investigate whether intron amplification products could, at least, distinguish congener species of agricultural importance within these orders. These included two commercial *Raphanus sativus* mitotypes confirmed by PCR markers as described by Kim et al. (2007), representatives of the two major *Phaseolus vulgaris* gene pools (Bhakta et al., 2017), congener species representatives of *Cenchrus, Citrus*, *Cynodon*, *Solanum* and *Vaccinium*, along with *Poncirus trifoliata*, which is cross-compatible with *Citrus* species (Moreira et al., 2002) and considered by some to fall within the genus *Citrus* (Ollitrault et al., 2020). *Citrus* materials were from the University of Florida Citrus Research and Education Center, Lake Alfred, Florida and Harris Citrus Nursery, Lithia, FL. The *Cynodon* entries were from the USDA National Plant Germplasm System. The *Phaseolus*, *Cenchrus*, *Solanum* and *Vaccinium* entries were obtained from the University of Florida research programs of Dr. C.E. Vallejos, Dr. L. Sollenberger, Dr. C.E. Vallejos, and Dr. J. Olmstead, respectively. Total cellular DNA was extracted from leaf samples by a modification of the cetyl trimethylammonium bromide (CTAB) method in which 50 mg of tissue was combined with 750 µl of CTAB buffer (Murray et al., 1980) and 10 µg of DNase-free RNase A in a FastPrep™ Lysing Matrix A tube, disrupted for 40 s in a FastPrep®-24 Instrument (MP Biomedicals LLC, Santa Ana, CA) and incubated at 65°C for 5 min. Cellular and lysing matrix debris was removed by centrifugation at 13,000 xg for 10 min at room temperature. Supernatants were extracted with an equal volume of chloroform-isoamyl alcohol mixed in a ratio of 24:1. DNA was precipitated from the aqueous phase by the addition of a 2/3 volume of isopropyl alcohol and recovered by centrifugation at 13,000 xg for 10 min at room temperature. The pellets were washed in 750 µl of 70% ethanol, air dried and rehydrated in 80 ul of 1 mM Trizma base, 0.1 mM di-sodium ethylene diamine tetra acetic acid (Na_2EDTA), 1 mM NaCl, pH 8. The concentration of DNA samples was determined from the absorbance at 260 nm.

DNA amplification and fractionation

The PCR primers used in this work (Table 2; Grosser, 2011) were designed against introns of the mitochondrial *nad1, nad2, nad4, nad5, nad7* and *cyctochrome c maturation Fc* (*ccmFc*) genes because these introns were common to seven plant species'

TABLE 1 Plant materials and intron sequence GenBank accession numbers.

Genus Species	Cultivar/Accession	GenBank Accession		
		*ccmFc*i1	*nad5*i4	*nad7*i1
Cenchrus amercianus[a]	TifLeaf3	OP800670	OP800688	OP800704
Cenchrus purpureus	Merkeron	OP800671	OP800689	OP800705
Citrus maxima	Hirado Buntan Pummelo	OP800658	OP800674	OP800690
Citrus japonica	Meiwa	OP800662	OP800678	OP800694
Citrus medica	Etrog	OP800661	OP800677	OP800693
Citrus paradisi[b]	Ruby Red	OP800659	OP800675	OP800691
Citrus reticulata	Ponkan	OP800660	OP800676	OP800692
Citrus sinensis[c]	Valencia	ND[d]	ND	ND
Cynodon dactylon	Royal Cape/PI290868	OP800668	OP800686	OP800702
Cynodon transvaalensis	Frankenwald Fine/PI290905	OP800669	OP800687	OP800703
Phaseolus vulgaris	Jamapa (Mesoamerican)	OP800673	OP800685	OP800701
Phaseolus vulgaris	Calima (Andean)	OP800672	OP800684	OP800700
Poncirus trifoliata[e]	English Large Flower	OP800663	OP800679	OP800695
Raphanus sativus	Red Velvet[f]	ND	ND	ND
Raphanus sativus	April Cross[g]	ND	ND	ND
Solanum lycopersicum	Bonny Best	OP800664	OP800680	OP800696
Solanum pennellii	LA716	OP800665	OP800681	OP800697
Vaccinium corymbosum	Bluecrop	OP800666	OP800682	OP800698
Vaccinium virgatum	Tifblue	OP800667	OP800683	OP800699

[a]*Cenchrus americanus* (*Pennisetum glaucum*, pearl millet) hybrid with wild *P. americanum* subsp. *Monodii* cytoplasm (Hanna, 1997; Hanna et al., 1997).
[b]*Citrus maxima* maternal lineage.
[c]*Citrus reticulata* maternal lineage.
[d]ND, sequence not determined.
[e]Considered by some as *Citrus trifoliata* (Ollitrault et al., 2020).
[f]F1 hybrid Harris Seeds 11701-00-00; commercial seed mixture or heteroplasmy prevented acquiring intron sequences.
[g]F1 Hybrid Harris Seeds 11700-00-01: commercial seed mixture or heteroplasmy prevented acquiring intron sequences.

mitochondrial genomes: *A. thaliana* (Unseld et al., 1997), *B. napus* (Handa, 2003), *B. vulgaris* (Kubo et al., 2000), *N. tabacum* (Sugiyama et al., 2005), *O. sativa* (Notsu et al., 2002), *T. aestivum* (Ogihara et al., 2005), and *Z. mays* (Allen et al., 2007). The National Center for Biotechnology Information (NCBI) accession numbers for these genomes are NC_001284, NC_002511, NC_008285, NC_006581, NC_007886, NC_007579, and NC_007982, respectively (https://www.ncbi.nlm.nih.gov/genome/organelle/, accessed 1/20/2023). Primer pairs were designed manually based upon the highly conserved coding regions flanking intron sequences or, in some cases, from conserved sequences within introns.

PCR amplification reactions were performed on replicate DNA preparations made from different plants of each entry, with the exception of the two *Cynodon* entries. For these only a single pot culture was available, so replicate DNA extractions were prepared from the single culture of each. PCR reactions of 50 µl contained 25-100 ng of DNA, 0.2 µM of each primer, 0.125 mM dNTPs, 1.25 units of high fidelity, TAKARA EXTAQ Hot Start DNA polymerase (Clontech, Mountain View, CA) in 1X TAKARA Hot Start reaction buffer. This high-fidelity polymerase was selected due to the length of the amplified introns and the intent to sequence PCR products. Amplification was for 30 cycles of 1 min at 94°C, 2 min at 55°C, and 3 min at 72°C. Electrophoresis through 1% agarose gels was performed to survey PCR reactions for successful amplification. The DNA Hyperladder II (Bioline Inc., Cambridge, MA) was used as a size marker. Electrophoresis was at 100V for 100 min in Tris-Borate-EDTA (TBE) buffer (10 mM Trizma base, 10 mM boric acid, 2.5 mM Na_2EDTA, pH 8.2). Gels were stained in 0.5 µg/ml ethidium bromide for 20 min and viewed over a UV transilluminator in a Molecular Imager® Gel Doc™ XR System (Bio Rad Laboratories, Inc. Hercules, CA). Gel images were captured with the Quantity One® 1-D Analysis Software (Bio Rad Laboratories, Inc.) and exported as.tif files. The AdvanCE™ FS96 capillary electrophoresis system (Advanced Analytical Technologies Inc., Ames, IA) was used to estimate the length of PCR amplification products in DNA base pairs (bp). Amplification products were diluted 1:15 in TE buffer (10 mM Trizma Base 1 mM Na_2EDTA,

TABLE 2 Primers for amplification and sequencing of plant mitochondrial introns.

Intron	Forward primer sequence (5′-3′)	Reverse primer sequence (5′-3′)
*ccmFc*i1	TTTCACATGGAGGAGTGTGC	TTCCCCATATGGAGTTCG
*ccmFc*i1	ATTGGTCAGACGACGACTACT[a]	TCTCTCAGTGTGGTCAGC[a]
*nad1*i2	CGATCTGCAGCTCAAATGGT	ACCTACAGCCCTTTCCTCT
*nad2*i1	GTAATGTGGGTTGGCTTGGA	GCAATAGTTAGGAGAGGTG
*nad2*i4	CAGTGGGAGTAGTGACTAG	GGAAGTCATTGCTAGTAG
*nad4*i1	AGGGGCCTTGTGCAGTAAA[b]	CTTTCTTTGTCTCGAACCCC
*nad4*i3	GTAGTACCGGTGAACCAGAT[b]	CTTACGGATGTATGCATG
*nad5*i1	ATGTTTGATGCTTCTTGGGG	TTAACATCACTACGGTCGGG
*nad5*i4	GGTATCTCGTACACATTCCG	CCCACATACGAGAAAAGGTC
*nad5*i4	CAACTAGTATAGTATAGCAG[a]	GGGAATCTAGGAATGAATGG[a]
*nad7*i1	AACGGAGAAGTGGTGGAACG	TTTCTCAGTCCCTCTAGTCG
*nad7*i1	AAGACCGTCTGGCGAAAACG[a]	CGTTTTCGCCAGACGGTCTT[a]
*nad7*i2	AGATGCCAGCGGAATGAT	GTGTTCTTGGGCCATCATAG
*nad7*i3	ATGTTAAGAGGTCGTGCG	AACATCGTAAGGTGCTGCTC

[a]Internal primer for intron sequencing.
[b]Primer binds near terminus but within the target intron.

pH 8) and fractionated by use of the DNF-915 dsDNA 915 Reagent Kit (Advanced Analytical Technologies Inc.) according to the supplier's instructions. Indel polymorphisms were confirmed by electrophoresis of DNA amplification products, individually and mixed, through Criterion™ precast 5% polyacrylamide gels (Bio-Rad Laboratories Inc., Hercules, CA) run in TBE buffer for 740 Volt-h and imaged as described above.

DNA sequencing and sequence analysis

The amplification products of *ccmFc*i1, *nad5*i4, *nad7*i1were purified for DNA sequencing through use of the QIAquick PCR Purification Kit (Qiagen Inc., Valencia, CA) according to supplier's instructions. Purified amplification products were fully sequenced in both directions by the University of Florida Interdisciplinary Center for Biotechnology Research (ICBR) Sanger Sequencing Core Laboratory in Gainesville, FL or by Eurofins USA. Intron sequences and their corresponding GenBank Accession numbers are listed in Table 1. The sequences were aligned on the MultAlin web server (Corpet, 1988) (http://multalin.toulouse.inra.fr/multalin/, accessed 9/2/2022). Nucleotide substitutions per site (K_0) were calculated by the formula of Kimura (1980) based upon pairwise alignments of sequences with all indels removed. Indels per site (I) were calculated as the number of indels in a pairwise alignment divided by the number of nucleotides in the alignment with indels removed (Laroche et al., 1997). Intron sequences found to differ between congener species were also analyzed for potential restriction fragment polymorphisms with the NEB cutter V 2.0 tool (Vincze et al., 2003) (http://nc2.neb.com/NEBcutter2/index.php, accessed 1/26/2023).

In silico prediction of intron amplification products

Prediction of intron amplification products across a wider range of plant taxa was performed through application of the Primer-BLAST tool (Ye et al., 2012) (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) to selected plant mitochondrial genomes in the NCBI organelle genome database (https://www.ncbi.nlm.nih.gov/genome/organelle/) (both accessed 1/20/2023). Genomes queried included early andiosperms *Magnolia biondii* (NC_049134.1) (Dong et al., 2020) and *Magnolia officinalis* (NC_064401) (unpublished), which could potentially differ in sequence from later diverged andiosperms. Additional orders of monocots were selected to complement the single order (Poales) investigated experimentally. These included *Allium cepa* male-sterilizing (KU318712.1) (Kim et al., 2016) and normal (AP018390.1) (Tsujimura and Terachi, 2018) cytoplasms representing monocot order Asparagales; *Cocos nucifera* (KX028885.1) (Aljohi et al., 2016) representing monocot order Arecacales; and *Zostera japonica* (NC_068803.1) (Chen et al., 2022) and *Zostera marina* (KX808392.1)(Petersen et al., 2017) representing monocot order Alismatales. Also included were dicots *Silene conica* (JF40490.1-JF50629.1), *Silene noctiflora* (KP053825.1-KP053880.1), *Silene latifolia* (HM562727.1) and *Silene vulgaris* (JF750427.1-JF750430.1). *Silene* is an important model genus that includes species exhibiting unusual patterns of mitochondrial genome expansion and nucleotide substitution, which potentially affect primer performance and utility. *Silene conica* and *Silene noctiflora* provide tests of primers on expanded mitochondrial genomes that exhibit accelerated nucleotide substitution rates in comparison to *Silene latifolia* and *Silene vulgaris* (Sloan et al., 2012).

Results

Mitochondrial intron amplification across angiosperm taxa

The intron primer sets (Table 2) successfully amplified the target intron in each of the 19 entries investigated (Figure 1, Table 3). PCR reactions generally produced a single major product, although additional products of low abundance were detected for some *nad2*i1, *nad5*i4 and *nad7*i3 amplifications (Figure 1).

The increasing number of complete plant mitochondrial genome sequences enabled investigation of the potential for these primer sets to amplify target introns in additional taxa. Primer-BLAST analysis of selected fully sequenced mitochondrial genomes predicted successful application of the introns in early angiosperms represented by *Magnolia biondii* and *Magnolia officionalis*; additional orders of monocots represented by *Allium cepa*, *Cocos nucifera*, *Zoster japonica* and *Zoster marina*; and an additional order of dicots represented by *Silene conica*, *Silene latifolia*, *Silene noctifolora* and *Silene vulgaris* (Table 4). Of the 121 primer-

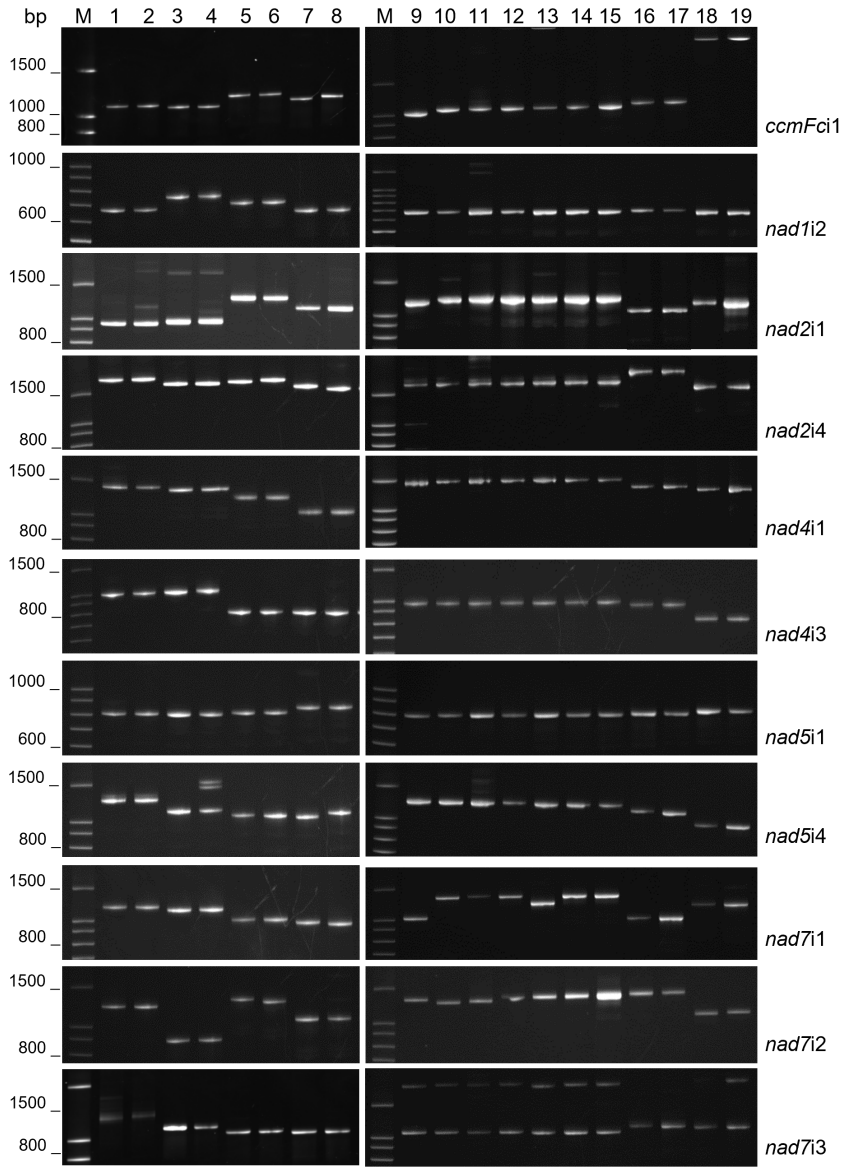


FIGURE 1
Mitochondrial intron lengths vary between but are conserved within plant genera. PCR amplification products of 11 plant mitochondrial introns were analyzed by polyacrylamide gel electrophoresis. M corresponds to a 100 base pair (bp) DNA ladder. DNA templates for PCR were as follows: 1) *Solanum pinnellii* LA716, 2) *Solanum lycopersicon* Bonny Best, 3) *Raphanus sativus* Red Velvet, 4) *Raphanus sativus* April Cross, 5) *Cynodon dactylon* Royal Cape, 6) *Cynodon transvaalensis* Frankenwald Fine, 7) *Cenchrus americanus* Tifleaf3, 8) *Cenchrus purpureus* Merkeron, 9) *Poncirus trifoliata* English Large Flower, 10) *Citrus japonica* Meiwa, 11) *Citrus medica* Etrog, 12) *Citrus maxima* Hirado Buntan, 13) *Citrus reticulata* Ponkan, 14) *Citrus paradisi* Ruby Red, 15) *Citrus sinensis* Valencia, 16) *Vaccinium virgatum* Tifblue, 17) *Vaccinium corymbosum* Blue Crop, 18) *Phaseolus vulgaris* Calima, 19) *Phaseolus vulgaris* Jamapa.

TABLE 3 Intron PCR product length[a] estimated by Advance™ capillary electrophoresis.

Entry	*ccmFc*i1	*nad1*i2	*nad2*i1	*nad2*i4	*nad4*i1	*nad4*i3	*nad5*i1	*nad5*i4	*nad7*i1	*nad7*i2	*nad7*i3
S. pennellii LA716[b]	1041	632	1133	1680	1436	751	900	1283	975	1295	1205
S. lycopersicum Bonny Best[b]	1048	644	1136	1682	1441	757	898	1290	955	1307	1205
P. vulgaris Calima[c]	4284	630	1408	1806	1458	739	926	1087	954	1212	1153
P. vulgaris Jamapa[c]	4312	638	1411	1806	1452	743	918	1090	963	1200	1160
V. corymbosum Blue Crop[d]	1072	649	1262	1858	1446	737	901	1302	866	1377	1139
V. virgatum Tifblue[d]	1067	644	1254	1847	1424	742	890	1299	865	1376	1142
R. sativus Red Velvet[e]	1064	673	1083	1896	1475	742	892	1154	1048	898	1134
R. sativus April Cross[e]	1061	678	1079	1899	1476	742	895	1167	1055	916	1132
C. dactylon PI290868[f]	1122	610	1375	1621	1278	676	926	1088	941	1109	1086
C. transvaalensis PI290695[f]	1116	606	1364	1558	1272	678	925	1099	933	1124	1080
C. americanus Tifleaf3[g]	1098	608	1358	1512	1009	675	912	1084	916	1134	1073
C. purpureus Merkeron[g]	1092	606	1363	1543	995	675	919	1030	937	1126	1076
P. trifoliata English Large[h]	1041	641	1262	1847	1468	758	886	1233	943	1421	1145
C. japonica Meiwa[h]	1085	643	1301	1842	1475	755	892	1238	977	1414	1158
C. medica Etrog[h]	1078	642	1274	1858	1484	755	890	1236	950	1418	1150
C. maxima Hirado Buntan[h]	1078	644	1302	1858	1477	755	893	1236	969	1424	1173
C. reticulata Ponkan[h]	1071	641	1322	1874	1485	755	892	1240	944	1433	1153
C. paradisi Ruby Red[h]	1077	648	1302	1853	1470	755	892	1235	959	1423	1157
C. sinensis Valencia[h]	1076	645	1298	1864	1475	753	893	1241	952	1437	1167
Range	1041 -4312	606 - 678	1083 -1411	1512 - 1899	995 - 1485	675 - 758	886 - 926	1030 - 1299	865 - 1055	898 - 1437	1073 - 1205

[a]PCR product sizes reported in DNA nucleotide pairs are the means of two biological replicates, or two technical replicates for *C. dactylon* and *C. transvaalensis*.
[b]Genus *Solanum* representing Eudicot order Solanales.
[c]Genus *Phaseolus* representing Eudicot order Fabales; Calima and Jamapa representing the Andean and Mesoamerican gene pools, respectively.
[d]Genus *Vaccinium* representing Eudicot order Ericales.
[e]Genus *Raphanus* representing Eudicot order Brassicales.
[f]Genus *Cynodon* representing Monocot order Poales.
[g]Genus *Cenchrus* representing Monocot order Poales.
[h]Genus *Poncirus* or *Citrus* representing Eudicot order Sapindales.

accession combinations tested, 79 predicted a single amplification product produced by perfectly matched primers. An additional 18 combinations predicted a single amplification product produced by primers with only one or two mis-matched nucleotides between the target genome and primer set. The 11 primer sets are therefore predicted to be useful for the amplification of mitochondrial introns across the angiosperms. Primers were predicted to be less effective for plant mitochondrial genomes that exhibit exceptionally high rates of genome expansion and nucleotide substitution. *Silene conica* and *Silene noctiflora* represent expanded mitochondrial genomes with accelerated nucleotide substitution rates in comparison to *Silene latifolia* and *Silene vulgaris* (Sloan et al., 2012). While all primer sets were predicted to amplify single products in *Silene latifolia* and *Silene vulgaris*, most primer sets predicted multiple, weak matches to *Silene conica* and *Silene noctiflora*. Nevertheless, 3-4 primer sets were still predicted to work well for these two templates (Table 4).

Intron length polymorphisms

The fractionation of experimentally produced intron amplification products by gel electrophoresis (Figure 1) and AdvanCE™ FS96 capillary electrophoresis (Table 3) demonstrated significant intron length variation among diverse angiosperm genera, in agreement with primer-BLAST observations (Table 4). Intron lengths, as estimated by the AdvanCE™ capillary technique, varied across genera by as few as 40 nucleotides in the case of *nad5*i1 to as many as 539 nucleotides in the case of *nad7*i2. This was excluding the extreme size (4284 nucleotides) of *Phaseolus ccmFc*i1, which likely reflects a split intron. Length polymorphisms between congener species were, however, few in number and challenging to detect by electrophoresis. The well-to-well variation of the AdvanCE™ FS96 precluded use of length values to detect small indel polymorphisms in relatively large DNA amplification products.

TABLE 4 Intron PCR product length predicted by Primer-BLAST[a].

Entry	*ccmFc*i1	*nad1*i2	*nad2*i1	*nad2*i4	*nad4*i1	*nad4*i3	*nad5*i1	*nad5*i4	*nad7*i1	*nad7*i2	*nad7*i3
Allium cepa[b] CMS-S	1142	588	1576	1621	1336	1988	903	1266	1382	958	1301
Allium cepa[b] Normal	1142	596	1576	1611	1336	1988	903	1266	1410	958	1301
Cocos nucifera[c]	1080	624	1313 2067[d]	1552	1358	2368	916	1072	924	1579	1056
Magnolia biondii[e]	1133 1112	626	1433	1569 1332[d]	1380	2391	891	1400	925	1532 And MWT[f]	1059
Magnolia officinalis[e]	1151	632	1451	1584 and MWT	1380	2437	901	1430	938 1490[d]	1563 433[d]	1079
Silene conica[g] isolate ABR	1055 1105[d]	MWT	1028	1273 3712[d]	3165[d]	1674	894	1162 2659[d]	1534 and MWT	565 1245[d]	837 2254[d]
Silene latifolia alba	1067	655	1131	1405	1500	1951	906	1136	1003	800	1173
Silene noctiflora[g] isolate BRP	NM[h]	MWT	1058 1123 2491[d]	MWT	1546 and MWT	707	920	1151	889	653 and MWT	NM[h]
Silene vulgaris isolate SD2	1067	672	1136	1409	1484	1975	924	1146	989	798	1169
Zostera japonica [i]	1621	552	1454	1808	1264	1456	993	1311	1283	788	2011
Zostera marina [i]	1937	549	1515 2640[d]	2128	1224	1456	999	1260	1283 1505[d]	788	2287

[a]PCR product sizes in DNA nucleotide pairs were predicted by NCBI Primer BLAST < https://www.ncbi.nlm.nih.gov/tools/primer-blast/> (accessed 1/23/2023). Numbers without superscripts indicate predicted PCR products with primers having 0-2 mismatches per primer on the target template.
[b]*Allium cepa* male-sterilizing (KU318712.1) and normal (AP018390.1) cytoplasms representing monocot order Asparagales with the male sterilizing cytoplasm of inter-specific origin (Manjunathagowda et al., 2021).
[c]*Cocos nucifera* (KX028885.1) representing monocot order Arecacales.
[d]Single weak target with 4 or 5 template mismatches per primer.
[e]*Magnolia bondii* (NC_049134.1) and *Magnolia officinalis* (NC_064401) representing early angiosperm Magnoliales.
[f]MWT, multiple weak targets with 4-5 template mismatches per primer.
[g]*Silene conica* (JF40490.1-JF50629.1) and *Silene noctiflora* (KP053825.1-KP053880.1) exhibit expanded genomes and accelerated nucleotide substitution rates in comparison to Silene latifolia (HM562727.1) and Silene vulgaris (JF750427.1-JF750430.1) (Sloan et al., 2012).
[h]No match to template.
[i]*Zostera japonica* (NC_068803.1) and *Zostera marina* (KX808392.1) representing monocot order Alismatales.

Length polymorphisms were identified by fractionation of amplification products on polyacrylamide gels (Figure 1) and confirmed by acrylamide gel electrophoresis of PCR product mixtures (Figure 2) for congeners of *Cenchrus* (*ccmFc*i1 and *nad2*i4), *Cynodon* (*nad7*i2) and *Citrus* (*nad7*i1 and *nad7*i2) species. Intron length polymorphisms are summarized in Table 5. The three *Citrus* maternal lineages and *C. japonica* were individually distinguished by the combination of *nad7*i1 and *nad7*i2 polymorphisms. *C. paradisi* (grapefruit) and *C. sinensis* (orange) were not distinguished from their respective *C. maxima* and *C. reticulata* maternal lineages. *Citrus* species were distinguished from *P. trifoliata* by length polymorphisms in *ccmFc*i1, *nad2*i1, *nad7*i1, and *nad7*i2. Electrophoresis did not, however, distinguish the introns of the two *Vaccinium* or *Solanum* species, *Phaseolus* gene pools, or *Raphanus sativus* mitotypes.

Primer-BLAST demonstrated that short length polymorphisms often distinguish congener species' mitochondrial introns (Table 4). In *Allium*, introns differing by 8, 10 and 28 nucleotides distinguished the male sterilizing cytoplasm, derived by interspecific introgression (Manjunathagowda et al., 2021), from the normal cytoplasm. *Magnolia biondii* and *Magnolia officinalis* varied in seven introns with length differences ranging from 6 to 46 nucleotides. *Silene vulgaris* differed from *Silene latifolia* in nine introns having length variations ranging from 4-24 nucleotides. *Zoster japonica* and *Zoster marina* were polymorphic with respect to length in eight introns. While five of these differences ranged from 3-61 nucleotides, length polymorphisms of 316, 320 and 276 nucleotides were predicted for *ccmFc*i1, *nad2*i4 and *nad7*i3, respectively. DNA sequence information clearly allows detection of mitochondrial intron length polymorphisms that distinguish related plant species.

Intron sequence analysis

*CcmFc*i1, *nad5*i4 and *nad7*i1 introns amplified from 16 entries were sequenced to further characterize indels detected by electrophoresis and to search for additional indels, along with SNPs (Figures S1, S2, and S3, respectively). *Citrus sinensis* (sweet orange with the *C. reticulata* maternal lineage) was not included, and heteroplasmy or seed mixtures in the two commercial

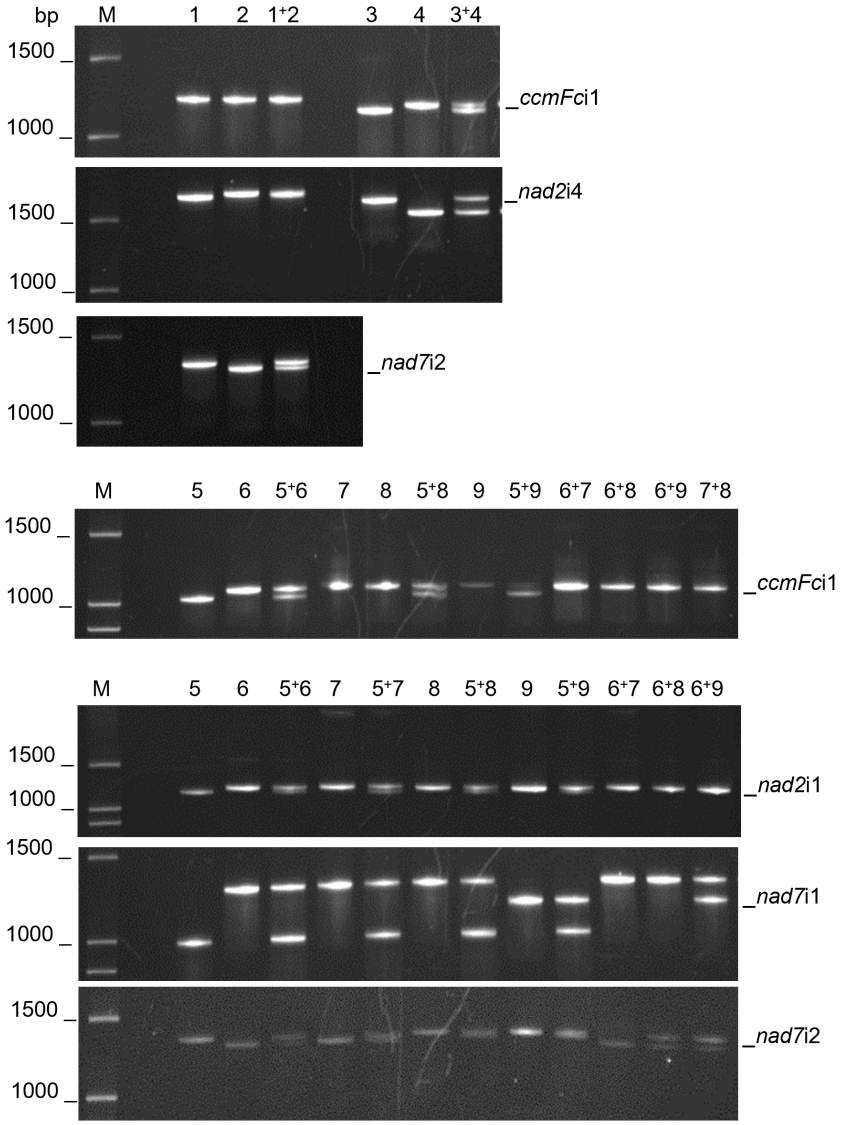


FIGURE 2
Mitochondrial intron length polymorphisms that distinguish related species. PCR amplification products of mitochondrial introns were separated by polyacrylamide gel electrophoresis. PCR products were analyzed individually and as mixtures to confirm the indel polymorphisms. M corresponds to a 100 base pair (bp) DNA ladder. DNA templates for PCR were as follows: 1) *Cynodon dactylon* Royal Cape, 2) *Cynodon transvaalensis* Frankenwald Fine, 3) *Cenchrus americanus* Tifleaf3, 4) *Cenchrus purpureus* Merkeron, 5) *Poncirus trifoliata* English Large Flower, 6) *Citrus japonica* Meiwa, 7) *Citrus medica* Etrog, 8) *Citrus maxima* Hirado Buntan, 9) *Citrus reticulata* Ponkan. Polymorphisms between *Cenchrus* spp. were confirmed for *ccmFc*i1 and *nad2*i4 and between *Cynodon* spp. for *nad7*i2. *CcmFc*i1, *nad2*i1, and *nad7*i1 polymorphisms differentiated *P. trifoliata* (5) from *Citrus* species (7-9). *Nad7*i1 also distinguished *C. reticulata* (9) from *C. japonica*, *C. medica* and *C. maxima* (6-8), whereas *nad7*i2 polymorphisms distinguished *P. trifoliata*, *C. maxima* and *C. reticulata* (5, 8, 9) from *C. japonica* and *C. medica* (6, 7).

Raphanus sativus accessions precluded obtaining quality sequences for comparison of these two mitotypes within this species. With respect to intra-species variation, *nad5*i4 and *nad7*i1 sequences did not distinguish the two gene pools of *Phaseolus vulgaris*. The *Phaseolus ccmFc*i1 shared 632 5' nucleotides and 133 3' nucleotides with other species separated by a 3353 nucleotide insertion (Figure S1B). The two *Phaseolus* accessions were polymorphic for one SNP and a 4 base indel within the 3353 nucleotide insertion, but were not polymorphic with respect to the intron regions. Moreover, the three *C. paradisi* introns were not polymorphic with respect to those of their *C. maxima* maternal ancestor. Sequencing further characterized indels detected by gel electrophoresis and revealed additional length polymorphisms (Table 5). The *ccmFc*i1 length polymorphism differentiating *P. trifoliata* from *Citrus* entries was due to separate deletions of 8, 9, and 17 nucleotides in *P. trifoliata* compared to *Citrus* (Figure S1A). Similarly, the polymorphism in *nad7*i1 was caused by separate deletions of 9, 8, and 9 nucleotides in *P. trifoliata* relative to *C. maxima, C. medica* and *C. japonica*. *C. reticulata* shared the 8 nucleotide deletion with *P. trifoliata*, while *C. medica* carried a unique 8 nucleotide deletion (Figure S3). The *ccmFc*i1 sequence distinguishing *Cenchrus* congeners was a 4 nucleotide indel (Figure S1A). Additional length polymorphisms identified by sequencing included a 4 nucleotide *nad5*i4 indel that distinguished *Cynodon*

TABLE 5 Experimentally identified intron length polymorphisms between congener species.

Intron	Polymorphic taxa			
	Allele 1[a]	Allele 2	Allele 3	Allele 4
*CcmFc*i1	*Cenchrus purpureus* (1008)	*Cenchrus americanus* (1004)		
*CcmFc*i1	*Citrus* ssp[b] (955)	*Poncirus trifoliata* (921)		
*nad2*i1	*Poncirus trifoliata*	*Citrus* ssp[b]		
*nad2*i4	*Cenchrus americanus*	*Cenchrus purpureus*		
*nad5*i4	*Cynodon dactylon* (929)	*Cynodon transvaalensis* (925)		
*nad7*i1	*Citrus maxima* (901) *Citrus japonica* (901)	*Citrus medica* (893)[c]	*Citrus reticulata* (893)[c]	*Poncirus trifoliata* (875)
*nad7*i2	*Cynodon dactylon*	*Cynodon transvaalensis*		
*nad7*i2	*Citrus maxima* *Citrus reticulata* *Poncirus trifoliata*	*Citrus medica* *Citrus japonica*		

[a]Allele 1 is designated the longest allele. (Intron lengths in nucleotides are indicated for those introns that were sequenced.).
[b]*Citrus japonica, Citrus medica, Citrus maxima, Citrus reticulata.*
[c]These entries carried different indels of the same length.

congeners (Figure S2) and a 4 nucleotide *nad7*i1 that distinguished *Cenchrus* congeners (Figure S3). Sequencing did not reveal indel polymorphisms between *Vaccinium* or *Solanum* congeners.

Sequences of three introns identified only nine SNPs that distinguished congener species (Table 6). The *nad5*i4 sequence alignment (Figure S2) revealed a SNP that distinguished *V. corymbosum* from *V. virgatum*. This was the only *Vaccinium* polymorphism identified in this study. Three *nad5*i4 SNPs distinguished *C. dactylon* from *C. transvaalensis* (Figure S2). In addition to the *nad7*i1 indels, a *nad7*i1 SNP was found to distinguish *C. reticulata* from other *Citrus* species (Figure S3). Of the nine SNPs, only one was a C/T difference that could possibly be erased at the RNA level by plant mitochondrial C-to-T RNA editing. In these comparisons, the frequency of SNPs per site (K_0) within genera was low - zero in the case of *ccmFc*i1. The average K_0 for *nad5*i4 and *nad7*i1 in congeneric species comparisons was 0.03 and 0.01, respectively, that of comparisons among dicot genera (Table 7). The frequency of indels per site (I) within genera was also low, 0.02-0.10 of I for comparisons among dicot genera (Table 7).

While sequence analysis is the most direct means of identifying length and SNP variation in amplified introns, these polymorphisms also create restriction pattern differences. Analysis of sequenced introns with NEB Cutter (Table S1) associated unique restriction patterns with the variant alleles reported in Tables 5 and 6. The only exception was the SNP that distinguished *Vaccinium corymbosum* and *Vaccinium virgatum nad5*i4 created no RFLPs across the 112 enzymes predicted by the NEB Cutter tool to cut these templates.

TABLE 6 Intron nucleotide[a] polymorphisms between congener species.

Intron	Polymorphic taxa		
	Allele 1	Allele 2	Allele 3
*nad5*i4	*Citrus maxima* 753 A 857 G	*Citrus reticulata*[b] 753 A 857 T	*Citrus medica* 753 C 857 G
*nad5*i4	*Vaccinium corymbosum* 982 G	*Vaccinium virgatum* 982 T	
*nad5*i4	*Cynodon dactylon* 659 G 660 T 662 T	*Cynodon transvaalensis* 659 T 660 C 662 A	
*nad7*i1	*Citrus reticulata* 787 C	*Citrus* ssp[c] 787 A	
*nad7*i1	*Cenchrus americanus* 764 A	*Cenchrus purpureus* 764 G	

[a]Nucleotides are numbered according to the multitaxa alignments shown in Figures S1-S3.
[b]Also *Citrus japonica* and *Poncirus trifoliata.*
[c]*Citrus maxima, Citrus medica Citrus japonica*, also *Poncirus trifoliata.*

TABLE 7 Average nucleotide substitutions (K_0) and indels (I) per site within genera and between dicot genera[b].

Intron	Within genera[a]		Between dicot genera[b]	
	K_0	I	K_0	I
*ccmFc*i1	0	0.0001 ± 0.0004	0.036 ± 0.008	0.005 ± 0.003
*nad5*i4	0.0012 ± 0.0009	0.0002 ± 0.0004	0.046 ± 0.006	0.008 ± 0.002
*nad7*i1	0.0002 ± 0.0004	0.0008 ± 0.0009	0.028 ± 0.003	0.007 ± 0.001

[a]Mean values ± standard deviation calculated for seven pair-wise species comparisons: *Citrus maxima - Citrus reticulata, Citrus maxima - Citrus medica, Citrus medica - Citrus reticulata, Cynodon dactylon - Cynodon transvaalensis, Cenchrus americanus - Cenchrus purpureus, Solanum lycopersicum - Solanum pennellii, Vaccinium corymbosum - Vaccinium virgatum.*
[b]Mean values ± standard deviation calculated for pair-wise species comparisons: *Citrus maxima - Solanum lycopersicum, Citrus maxima - Vaccinium corymbosum*, and *Solanum lycopersicum - Vaccinium corymbosum.*

Discussion

Universal primers for amplification of plant mitochondrial introns

The 11 PCR primer sets used in this work demonstrated robust amplification of the target mitochondrial introns across 16 species representing eight plant genera and seven plant orders. Primer-BLAST analysis with these same primer sets predicted successful amplification of mitochondrial introns from early angiosperms and additional orders of monocots and dicots. This expands and improves the available universal primers for plant mitochondrial introns (Demesure et al., 1995; Dumolin-Lapegue et al., 1997; Duminil et al., 2002). Aleksić (2016) found limited applicability of previously developed universal mitochondrial primers to legume (*Fabaceae*) species and suggested family-specific primers as a more practical approach. The primer sets employed here successfully amplified the mitochondrial introns of *P. vulgaris* as a representative legume. Previous universal primer design strategies (Duminil et al., 2002; Froelicher et al., 2011) utilized mitochondrial sequences conserved between *A. thaliana* and *B. vulgaris* only. Primer design based on conserved introns and flanking sequences from seven plant species (Grosser, 2011) likely contributed to the extended applicability of the current primer sets. Although most plant species' mitochondrial genomes evolve slowly with respect to coding sequences (Wolfe et al., 1987; Palmer and Herbon, 1988), plant genera containing taxa with widely varying rates of mitochondrial nucleotide substitution have been identified (Cho et al., 2004; Parkinson et al., 2005; Mower et al., 2007; Sloan et al., 2009). Primer-BLAST analysis did predict that some, but not all, of the 11 primer sets would work reliably on the *Silene* species having rapidly evolving mitochondrial coding sequences. A further complication with *Silene conica* and *Silene noctiflora* is that their highly expanded genomes apparently contain multiple, degenerate targets for the intron flanking primers (Table 4).

Intron polymorphism between and within genera

Mitochondrial intron length polymorphisms detectable by electrophoretic techniques were frequently observed between genera, whereas comparisons within genera revealed primarily short intron length variations. Large indels are therefore tolerated within introns, but rates of such variation are low within genera. These contrasting observations likely reflect the evolutionary processes that shaped modern plant organellar group II introns from their self-splicing, progenitor introns. On the one hand, altered intron sequences combined with novel nuclear and organelle-encoded splicing factors to maintain competence for splicing while shifting away from the group II ribozymic, self-splicing structures (Bonen, 2008; Brown et al., 2014). At the same time, the requirement for splicing factors to evolve in concert with the intron structure likely constrained variants that can be successfully spliced (de Longevialle et al., 2010; Zimmerly and Semper, 2015). Plant organelle introns retain significant common structural features (Bonen, 2008). Moreover, they reside within genes essential to photosynthesis or respiration, creating selective pressure for the maintenance of efficient splicing (Brown et al., 2014; Zimmerly and Semper, 2015; Best et al., 2020). Arrays of protein factors are required for the splicing of plastid and mitochondrial introns. These include members of the maturase family, descended from the maturases encoded in ribozymic, self-splicing group II introns (Schmitz-Linneweber et al., 2015), along with APO, CRM, PORR, PPR and TERF families of RNA binding proteins. With the exception of one plastid and one mitochondria-encoded maturase, these proteins are encoded by the nuclear genome and imported into the organelles where they act in combinatorial fashion for the splicing of particular introns or groups of introns (de Longevialle et al., 2010; Brown et al., 2014; Zimmerly and Semper, 2015; Wang et al., 2022). The complexity and specificity of this process may explain the lack of large intron indels found within genera.

The plant mitochondrial intron length differences between congener species as predicted by by Primer-BLAST averaged 17 nucleotides, excepting the three *Zoster* introns with larger differences. The experimentally characterized indels that distinguished congeners or cross-compatible species averaged less than 10 nucleotides in length, necessitating high-resolution acrylamide gels or DNA sequencing for discernment. Gel-resolved intron length polymorphisms differentiated *Citrus*, *Cenchrus*, and *Cynodon* congeners, but intron sequencing provided a more accurate picture of indel polymorphisms. The *nad7*i1 amplicons of *C. medica* and *C. reticulata*, for example,

carried different indels of the same length. Even when larger intron size differences were apparent within genera, sequencing revealed them to result from multiple short indels. This is consistent with prior reports that short indels (1-10 bp) comprise greater than 50% of indels in plant mitochondrial introns and probably originate from slipped strand mispairing events during replication (Laroche et al., 1997).

For each of the three introns sequenced in the present study, the average frequency of nucleotide substitutions per site (K_0) was also low within genera, but SNPs that distinguished congeners of *Citrus*, *Cynodon*, *Cenchrus*, and *Vaccinium* were identified. These can serve as useful markers through workflows such as cleaved amplified polymorphic sequence (CAPS), PCR combined with sequencing, or amplification refractory mutation analysis strategies (Lo, 1998; Ciarmiello et al., 2013). With one exception, SNPs and short indels that distinguished congeners' introns also created CAPS markers (Table S1). More broadly, K_0 values for comparisons between *Citrus*, *Vaccinium* and *Solanum* as representative dicot genera (Table 7) were similar to those reported by Laroche et al. (1997) for comparisons of six mitochondrial introns between two to three dicot genera. In the present study, K_0 values for intron sequence comparisons between congeneric species were 0.01-0.03 times those for comparisons between dicot genera. The low nucleotide substitution rates likely result from the low frequency of nucleotide substitutions characteristic of most plant mitochondrial genomes, typically three to ten times lower than nuclear nucleotide substitution rates (Wolfe et al., 1987; Palmer and Herbon, 1988; Drouin et al., 2008).

The limited variation of mitochondrial introns within plant genera contrasts with the extensive diversity of nuclear introns, which show a high frequency of length and substitution polymorphisms within species of *Oryza* (Wang et al., 2005), *Solanum* (Wang et al., 2010), *Allium* (Jayaswal et al., 2019) and *Medicago* (Shilpa and Lohithaswa, 2021) among others. While no mitochondrial intron polymorphisms distinguished *Solanum lycopersicum* from *S. pennellii*, two studies document extensive nuclear intron polymorphisms within *Solanum lycopersicum* (Van Deynze et al., 2007; Wang et al., 2010). Organellar group II introns are considered the ancestors of nuclear introns, which lack folding constraints because they share the use of spliceosomal RNAs that have taken on the functions of the group II intron domains (Sharp, 1991). The spliceosome is a complex that is highly malleable in order to accommodate diverse exon ends and alternative splicing, perhaps permitting more varied intron sequences (Chen and Moore, 2014).

Application of plant mitochondrial intron polymorphisms

When present, organelle intron polymorphisms have valuable applications for determining inheritance in sexual crosses or somatic hybridizations. The markers investigated in this study have proved useful for determining organelle inheritance in *Citrus* cybrids. Cybrids are produced spontaneously as a by-product of protoplast fusion and are characterized by the diploid nuclear genome of the mesophyll fusion partner, the mitochondrial genome of the embryogenic callus partner, and random inheritance of chloroplast DNA (Grosser et al., 1996; Cabasson et al., 2001; Guo et al., 2004; Guo et al., 2013). This contrasts with typical protoplast fusion products, which possess tetraploid nuclei inherited from both parents. Cybrids provide a means to quickly create novel combinations of nuclear and organellar genotypes and to evaluate their phenotypic consequences. Specific organelle genotypes are associated with beneficial traits in cybrids. For example, grapefruit cybrids with mandarin mitochondrial DNA exhibit an extended season of high-quality fruit (Satpute et al., 2015), whereas grapefruit cybrids with kumquat plastid DNA exhibit increased resistance to citrus canker regardless of mitochondrial origin (Omar et al., 2017). The *nad7*i1 and *nad7*i2 primer sets (Table 2) were utilized, respectively, for verification and characterization of mitochondrial DNA inheritance in these two sets of cybrids, illustrating application for mitochondrial intron markers.

The currently reported *nad7*i1 marker overlaps with and confirms one of the three markers that Froelicher et al. (2011) demonstrated to be polymorphic in *Citrus*. Because our primer set amplified the entire intron, a new SNP was added to the previously published indels. Moreover, the list of intron markers polymorphic for *Citrus* species was expanded to include *nad7*i2, *ccmFc*, and *nad2*i1. The additional polymorphic markers did not, however, further distinguish differences within the seven citrus mitotypes identified by Froelicher et al. (2011). For example, *C. maxima* and its maternal derivative *C. paradisi* remained indistinguishable for all introns sequenced in this study.

Due to the lack of conserved gene order among plant mitochondrial genomes, often even between closely related taxa, assembling plant mitochondrial genome sequences presents special challenges and can preclude the universal application of intergenic sequences for distinguishing between closely related groups (Duminil and Besnard, 2021). Mitochondrial intron markers have demonstrated applicability in studies of population genetics, genotype characterization, detection of past hybridizations, and biogeographic studies of gene pool distributions (Ciarmiello et al., 2013; Aizawa et al., 2014; Xiang et al., 2014; Kersten et al., 2015). The current study documents a widely applicable set of primers for the mitochondrial marker toolbox and provides insights into the conservation and variation of plant mitochondrial introns.

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: https://www.ncbi.nlm.nih.gov/genbank/, OP800658-OP800705.

Author contributions

CC, FG, JGro, and JGra: designed the study and collected the genetic materials. KC, KL, MG, SS, MM and YL: conducted the research. CC and MG: wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was funded by the University of Florida Institute of Food and Agricultural Science. MG and SS were supported by the University of Florida University Scholars Program. MMM was supported by the Hunt Brothers Fellowship and the Ciencia sem Fronteiras Award 1245/13-9.

Acknowledgments

We thank Drs. Jim Olmstead, Lynn Sollenberger and Eduardo Vallejos for sharing genetic materials used in this study.

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.2023.1116851/full#supplementary-material

References

Aizawa, M., Yoshimaru, H., Takahashi, M., Kawahara, T., Sugita, H., Saito, H., et al. (2014). Genetic structure of Sakhalin spruce (*Picea glehnii*) in northern Japan and adjacent regions revealed by nuclear microsatellites and mitochondrial gene sequences. *J. Plant Res.* 128, 91–102. doi: 10.1007/s10265-014-0682-7

Aleksić, J. M. (2016). Family-specific vs. universal PCR primers for the study of mitochondrial DNA in plants. *Genetika* 48, 777–798. doi: 10.2298/GENSR1602777A

Aljohi, H. A., Liu, W., Lin, Q., Zhao, Y., Zeng, J., Alamer, A., et al. (2016). Complete sequence and analysis of coconut palm (*Cocos nucifera*) mitochondrial genome. *PloS One* 10. doi: 10.1371/journal.pone.0163990

Allen, J. O., Fauron, C. M., Minx, P., Roark, L., Oddiraju, S., Lin, G. N., et al. (2007). Comparisons among two fertile and three male-sterile mitochondrial genomes of maize. *Genetics* 177, 1173–1192. doi: 10.1534/genetics.107.073312

Archibald, J. M., and Richards, T. A. (2010). Gene transfer: anything goes in plant mitochondria. *BMC Biol.* 8, 147. doi: 10.1186/1741-7007-8-147

Bastien, D., Favre, J. M., Collignon, A. M., Sperisen, C., and Jeandroz, S. (2003). Characterization of a mosaic minisatellite locus in the mitochondrial DNA of Norway spruce [*Picea abies* (L.) karst.]. *Theor. Appl. Genet.* 107, 574–580. doi: 10.1007/s00122-003-1284-2

Besse, P. (2021). "Guidelines for the choice of sequences for molecular plant taxonomy," in *Molecular plant taxonomy. methods in molecular biology*, vol. vol. 2222 . Ed. P. Besse (New York: Humana). doi: 10.1007/978-1-0716-0997-2_2

Best, C., Mizrahi, R., and Ostersetzer-Biran, O. (2020). Why so complex? the intricacy of genome structure and gene expression, associated with angiosperm mitochondria may relate to the regulation of embryo quiescence or dormancy-intrinsic blocks to early plant life. *Plants* 9, 598. doi: 10.3390/plants9050598

Bhakta, M. S., Gezan, S. A., Michelangeli, C. J. A., Carvalho, M., Zhang, L., Jones, J. W., et al. (2017). A predictive model for time-to-flowering in the common bean based on QTL and environmental variables. *G3: Genes Genomes Genet.* 7, 3901–3912. doi: 10.1534/g3.117.300229

Bock, D. G., Andrew, R. L., and Rieseberg, L. H. (2014). On the adaptive value of cytoplasmic genomes in plants. *Mol. Ecol.* 20, 4899–4911. doi: 10.1111/mec.12920

Bonen, L. (2008). Cis-and trans-splicing of group II introns in plant mitochondria. *Mitochondrion* 8, 26–34. doi: 10.1016/j.mito.2007.09.005

Brown, G. G., des Francs-Small, C. C., and Ostersetzer-Biran, O. (2014). Group II intron splicing factors in plant mitochondria. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00035

Cabasson, C. M., Luro, F., Ollitrault, P., and Grosser, J. W. (2001). Non-random inheritance of mitochondrial genomes in *Citrus* hybrids produced by protoplast fusion. *Plant Cell Rep.* 20, 604–609. doi: 10.1007/s002990100370

Camus, M. F., Alexander-Lawrie, B., Sharbrough, J., and Hurst, G. D. D. (2022). Inheritance through the cytoplasm. *Heredity* 129, 31–43. doi: 10.1038/s41437-022-00540-2

CBOL Plant Working Group (2009). A DNA barcode for land plants. *Proc. Natl. Acad. Sci. U. S. A.* 106, 12794–127987. doi: 10.1073/pnas.0905845106

Chandra, A., Jain, R., Solomon, S., Shrivastava, S., and Roy, A. K. (2013). Exploiting EST databases for the development and characterisation of 3425 gene-tagged CISP markers in biofuel crop sugarcane and their transferability in cereals and orphan tropical grasses. *BMC Res. Notes* 6, 47. doi: 10.1186/1756-0500-6-47

Chen, W., and Moore, M. J. (2014). The spliceosome: Disorder and dynamics defined. *Curr. Opin. Struct. Biol.* 24, 141–149. doi: 10.1016/j.sbi.2014.01.009

Chen, J., Zang, Y., Liang, S., Xue, S., Shang, S., Zhu, M., et al. (2022). Comparative analysis of mitochondrial genomes reveals marine adaptation in seagrasses. *BMC Genom* 23. doi: 10.1186/s12864-022-09046-x

Cho, Y., Mower, J. P., Qiu, Y. L., and Palmer, J. D. (2004). Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. *Proc. Natl. Acad. Sci. U S A.* 101, 17741–17746. doi: 10.1073/pnas.0408302101

Ciarmiello, L. F., Pontecorvo, G., Piccirillo, P., De Luca, A., Carillo, P., Kafantaris, I., et al. (2013). Use of nuclear and mitochondrial single nucleotide polymorphisms to characterize English walnut (*Juglans regia* l.) genotypes. *Plant Mol. Biol. Rep.* 31, 1116–1130. doi: 10.1007/s11105-013-0575-2

Colombatti, F., Gonzalez, D. H., and Welchen, E. (2014). Plant mitochondria under pathogen attack: a sigh of relief or a last breath? *Mitochondrion* 19 Pt B, 238–244. doi: 10.1016/j.mito.2014.03.006

Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucl. Acids Res.* 16, 10881–10890. doi: 10.1093/nar/16.22.10881

de Freitas, K. E. J., Busanello, C., Viana, V. E., Pegoraro, C., de Carvalho, V. F., da Maia, L. C., et al. (2022). An empirical analysis of mtSSRs: Could microsatellite distribution patterns explain the evolution of mitogenomes in plants? *Funct. Integr. Genomics* 22, 35–53. doi: 10.1007/s10142-021-00815-7

de Longevialle, A. F., Small, I. D., and Lurin, C. (2010). Nuclearly encoded splicing factors implicated in RNA splicing in higher plant organelles. *Mol. Plant* 3, 691–705. doi: 10.1093/mp/ssq025

Demesure, B., Sodzi, N., and Petit, R. J. (1995). A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol. Ecol.* 4, 129–134. doi: 10.1111/j.1365-294x.1995.tb00201.x

Dong, S., Chen, L., Liu, Y., Wang, Y., Zhang, S., Yang, L., et al. (2020). The draft mitochondrial genome of magnolia biondii and mitochondrial phylogenomics of angiosperms. *PloS One* 15. doi: 10.1371/journal.pone.0231020

Dourmap, C., Roque, S., Morin, A., Caubrière, D., Kerdiles, M., Béguin, K., et al. (2020). Stress signalling dynamics of the mitochondrial electron transport chain and oxidative phosphorylation system in higher plants. *Ann. Bot.* 125, 721–736. doi: 10.1093/aob/mcz184

Drouin, G., Daoud, H., and Xia, J. (2008). Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. *Mol. Phylogenet. Evol.* 49, 137–141. doi: 10.1016/j.ympev.2008.09.009

Duminil, J., and Besnard, G. (2021). "Utility of the mitochondrial genome in plant taxonomic studies," in *Molecular plant taxonomy. methods mol. biol*, vol. 2222 . Ed. P. Besse (New York: Humana). doi: 10.1007/978-1-0716-0997-2_6

Duminil, J., Pemonge, M. H., and Petit, R. J. (2002). A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial DNA. *Mol. Ecol. Notes* 2, 428–430. doi: 10.1046/j.1471-8286.2002.00263.x

Dumolin-Lapegue, S., Pemonge, M. H., and Petit, R. J. (1997). An enlarged set of consensus primers for the study of organelle DNA in plants. *Mol. Ecol.* 6, 393–397. doi: 10.1046/j.1365-294X.1997.00193.x

Egan, A. N., Schlueter, J., and Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology. *Am. J. Bot.* 99, 175–185. doi: 10.3732/ajb.1200020

Froelicher, Y., Mouhaya, W., Bassene, J. B., Costantino, G., Kamiri, M., Luro, F., et al. (2011). New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. *Tree Genet. Genomes* 7, 49–61. doi: 10.1007/s11295-010-0314-x

Galeano, C. H., Cortés, A. J., Fernández, A. C., Soler, Á., Franco-Herrera, N., Makunde, G., et al. (2012). Gene-based single nucleotide polymorphism markers for genetic and association mapping in common bean. *BMC Genet.* 13, 48. doi: 10.1186/1471-2156-13-48

Godbout, J., Jaramillo-Correa, J. P., Beaulieu, J., and Bousquet, J. (2005). A mitochondrial DNA minisatellite reveals the postglacial history of jack pine (*Pinus banksiana*), a broad-range north American conifer. *Mol. Ecol.* 14, 3497–3512. doi: 10.1111/j.1365-294X.2005.02674.x

Grosser, M. (2011) *Plant mitochondrial introns as genetic markers* (Gainesville, Florida: University of Florida). Available at: https://ufdc.ufl.edu/AA00060090/00001/pdf (Accessed 12/1/2022). undergraduate thesis.

Grosser, J. W., Gmitter, F. G., Tusa, N., Recupero, G. R., and Cucinotta, P. (1996). Further evidence of a cybridization requirement for plant regeneration from citrus leaf protoplasts following somatic fusion. *Plant Cell Rep.* 15, 672–676. doi: 10.1007/BF00231922

Gualberto, J. M., and Newton, K. J. (2017). Plant mitochondrial genomes: Dynamics and mechanisms of mutation. *Annu. Rev. Plant Biol.* 68, 225–252. doi: 10.1146/annurev-arplant-043015-112232

Guo, W. W., Prasad, D., Cheng, Y. J., Serrano, P., Deng, X. X., and Grosser, J. W. (2004). Targeted cybridization in citrus: Transfer of Satsuma cytoplasm to seedy cultivars for potential seedlessness. *Plant Cell Rep.* 22, 752–758. doi: 10.1007/s00299-003-0747-x

Guo, W. W., Xiao, S. X., and Deng, X. X. (2013). Somatic cybrid production *via* protoplast fusion for citrus improvement. *Sci. Hortic. (Amsterdam).* 163, 20–26. doi: 10.1016/j.scienta.2013.07.018

Gupta, S., Kumari, K., Das, J., Lata, C., Puranik, S., and Prasad, M. (2011). Development and utilization of novel intron length polymorphic markers in foxtail millet (*Setaria italica* (L.) p. beauv.). *Genome* 54, 586–602. doi: 10.1139/g11-020

Handa, H. (2003). The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* l.): comparative analysis of the mitochondrial genomes of rapeseed and a*rabidopsis thaliana*. *Nucl. Acids Res.* 31, 5907–5916. doi: 10.1093/nar/gkg795

Hanna, W. W. (1997). Registration of tift 8593 pearl millet genetic stock. *Crop Sci.* 37, 1412. doi: 10.2135/cropsci1997.0011183X003700040100x

Hanna, W. W., Hill, G. M., Gates, R. N., Wilson, J. P., and Burton, G. W. (1997). Registration of 'Tifleaf 3' pearl millet. *Crop Sci.* 37, 1388. doi: 10.2135/cropsci1997.0011183X003700040075x

Hodel, R. G. J., Segovia-Salcedo, C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., et al. (2016). The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Appl. Plant Sci.* 41. doi: 10.3732/apps.1600025

Honma, Y., Yoshida, Y., Terachi, T., Toriyama, K., Mikami, T., and Kubo, T. (2011). Polymorphic minisatellites in the mitochondrial DNAs of *Oryza* and *Brassica*. *Curr. Genet.* 57, 261–270. doi: 10.1007/s00294-011-0345-3

Hu, J., Huang, W., Huang, Q., Qin, X., Yu, C., Wang, L., et al. (2014). Mitochondria and cytoplasmic male sterility in plants. *Mitochondrion* 19 Pt B, 282–288. doi: 10.1016/j.mito.2014.02.008

Jaramillo-Correa, J. P., Aguirre-Planter, E., Eguiarte, L. E., Khasa, D. P., and Bousquet, J. (2013). Evolution of an ancient microsatellite hotspot in the conifer mitochondrial genome and comparison with other plants. *J. Mol. Evol.* 76, 146–157. doi: 10.1007/s00239-013-9547-2

Jayaswal, K., Sharma, H., Bhandawat, A., Sagar, R., Yadav, V. K., Sharma, V., et al. (2019). Development of intron length polymorphic (ILP) markers in onion (*Allium cepa* l.), and their cross-species transferability in garlic (*A. sativum* l.) and wild relatives. *Genet. Resour. Crop Evol.* 66, 1379–1388. doi: 10.1007/s10722-019-00808-3

Keeling, P. J. (2009). Role of horizontal gene transfer in the evolution of photosynthetic eukaryotes and their plastids. *Methods Mol. Biol.* 532, 501–515. doi: 10.1007/978-1-60327-853-9_29

Kersten, B., Voss, M. M., and Fladung, M. (2015). Development of mitochondrial SNP markers in different *Populus* species. *Trees* 29, 575–582. doi: 10.1007/s00468-014-1136-5

Kim, B., Kim, K., Yang, T. J., and Kim, S. (2016). Completion of the mitochondrial genome sequence of onion *(Allium cepa* l.) containing the CMS-s male-sterile cytoplasm and identification of an independent event of the ccmF n gene split. *Curr. Genet.* 62, 873–885. doi: 10.1007/s00294-016-0595-1

Kim, J. H., Lee, C., Hyung, D., Jo, Y. J., Park, J. S., Cook, D. R., et al. (2015). CSGM designer: A platform for designing cross-species intron-spanning genic markers linked with genome information of legumes. *Plant Methods* 11, 1–11. doi: 10.1186/s13007-015-0074-6

Kim, S., Lim, H., Park, S., Cho, K.-H., Sung, S.-K., Oh, D.-G., et al. (2007). Identification of a novel mitochondrial genome type and development of molecular markers for cytoplasm classification in radish (*Raphanus sativus* l.). *Theor. Appl. Genet.* 111, 1191–1200. doi: 10.1007/s00122-007-0639-5

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/BF01731581

Kubo, T., and Mikami, T. (2007). Organization and variation of angiosperm mitochondrial genome. *Physiol. Plant* 129, 6–13. doi: 10.1111/j.1399-3054.2006.00768.x

Kubo, T., Nishizawa, S., Sugawara, A., Itchodo, N., Estiati, A., and Mikami, T. (2000). The complete nucleotide sequence of the mitochondrial genome of sugar beet (*Beta vulgaris* l.) reveals a novel gene for tRNACys(GCA). *Nucl. Acids Res.* 28, 2571–2576. doi: 10.1093/nar/28.13.2571

Kumar, M., Kapil, A., and Shanker, A. (2014). MitoSatPlant: Mitochondrial microsatellites database of viridiplantae. *Mitochondrion* 19 Pt B, 334-337. doi: 10.1016/j.mito.2014.02.002

Laroche, J., Li, P., Maggia, L., and Bousquet, J. (1997). Molecular evolution of angiosperm mitochondrial introns and exons. *Proc. Natl. Acad. Sci. U. S. A.* 94, 5722–5727. doi: 10.1073/pnas.94.11.5722

Lessa, E. P. (1992). Rapid surveying of DNA sequence variation in natural populations. *Mol. Biol. Evol.* 9, 323–330. doi: 10.1093/oxfordjournals.molbev.a040723

Levings, C. S., and Pring, D. R. (1977). Diversity of mitochondrial genomes among normal cytoplasms of maize. *J. Hered.* 68, 350–354. doi: 10.1093/oxfordjournals.jhered.a108858

Li, C., Riethoven, J.-J. M., and Naylor, G. J. P. (2012). EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. *Mol. Ecol. Resour.* 12, 967–971. doi: 10.1111/j.1755-0998.2012.03167.x

Lo, Y. M. D. (1998). The amplification refractory mutation system. *Methods Mol. Med.* 16, 61–70. doi: 10.1385/0-89603-499-2:61. Clinical Applications of PCR.

Mahapatra, K., Banerjee, S., De, S., Mitra, M., Roy, P., and Roy, S. (2021). An insight into the mechanism of plant organelle genome maintenance and implications of organelle genome in crop improvement: An update. *Front. Cell Dev. Biol.* 10. doi: 10.3389/fcell.2021.671698

Manjunathagowda, D. C., Muthukumar, P., Gopal, J., Prakash, M., Bommesh, J. C., Nagesh, G. C., et al. (2021). Male Sterility in onion (*Allium cepa* l.): origin: origin, evolutionary status, and their prospectus. *Genet. Resour. Crop Evol.* 68, 421–439. doi: 10.1007/s10722-020-01077-1

Mitton, J. B., Kreiser, B. R., and Latta, R. G. (2000). Glacial refugia of limber pine (*Pinus flexilis* James) inferred from the population structure of mitochondrial DNA. *Mol. Ecol.* 9, 91–97. doi: 10.1046/j.1365-294X.2000.00840.x

Moore, G. A. (2001). Oranges and lemons: clues to the taxonomy of *Citrus* from molecular markers. *Trends Genet.* 17, 536–540. doi: 10.1016/s0168-9525(01)02442-8

Moreira, C. D., Gmitter, F. G.Jr., Grosser, J. W., Huang, S., Ortega, V. M., and Chase, C. D. (2002). Inheritance of organelle DNA sequences in a *Citrus-poncirus* intergeneric cross. *J. Hered.* 93, 174–178. doi: 10.1093/jhered/93.3.174

Mower, J. P., Touzet, P., Gummow, J. S., Delph, L. F., and Palmer, J. D. (2007). Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. *BMC Evol. Biol.* 7, 135. doi: 10.1186/1471-2148-7-135

Murray, M. G., Murray, W. F., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4325. doi: 10.1093/nar/8.19.4321

Nishizawa, S., Kubo, T., and Mikami, T. (2000). Variable number of tandem repeat loci in the mitochondrial genomes of beets. *Curr. Genet.* 37, 34–38. doi: 10.1007/s002940050005

Nishizawa, S., Mikami, T., and Kubo, T. (2007). Mitochondrial DNA phylogeny of cultivated and wild beets: Relationships among cytoplasmic male-sterility-inducing and nonsterilizing cytoplasms. *Genetics* 177, 1703–1712. doi: 10.1534/genetics.107.076380

Notsu, Y., Masood, S., Nishikawa, T., Nubo, K., Akiduki, G., Nakazono, M., et al. (2002). The complete sequence of the rice (*Oryza sativa* l.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. *Mol. Genet. Genom.* 268, 434–445. doi: 10.1007/s00438-002-0767-1

Ogihara, Y., Yamazaki, Y., Murai, K., Kanno, A., Terachi, T., Shiina, T., et al. (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. *Nucleic Acids Res.* 33, 6235–6250. doi: 10.1093/nar/gki925

Ollitrault, P., Curk, F., and Krueger, R. (2020). "Citrus taxonomy," in *The citrus genus*. Eds. M. Talon, M. Caruso and F. G. Gmitter Jr. (Amsterdam: Elsevier), 57–81. doi: 10.1016/B978-0-12-812163-4.00004-8

Omar, A. A., Murata, M., Yu, Q., Gmitter, F. G., Chase, C. D., Graham, J. H., et al. (2017). Production of three new grapefruit cybrids with potential for improved citrus canker resistance. *Vitr. Cell. Dev. Biol. - Plant* 53, 256–269. doi: 10.1007/s11627-017-9816-7

Palmer, J. D., and Herbon, L. A. (1988). Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. *J. Mol. Evol.* 28, 87–97. doi: 10.1007/BF02143500

Parkinson, C. L., Mower, J. P., Qiu, Y. L., Shirk, A. J., Song, K., Young, N. D., et al (2005). Multiple major increases and decreases in mitochondrial substitution rates in the plant family geraniaceae. *BMC Evol. Biol.* 5, 73. doi: 10.1186/1471-2148-5-73

Petersen, G., Cuenca, A., Zervas, A., Ross, G. T., Graham, S. W., Barrett, C. F., et al. (2017). Mitochondrial genome evolution in alismatales: Size reduction and extensive loss of ribosomal protein genes. *PloS One* 12. doi: 10.1371/journal.pone.0177606

Potter, K. M., Hipkins, V. D., Mahalovich, M. F., and Means, R. E. (2013). Mitochondrial DNA haplotype distribution patterns in *Pinus ponderosa* (Pinaceae): Range-wide evolutionary history and implications for conservation. *Am. J. Bot.* 100, 1562–1579. doi: 10.3732/ajb.1300039

Qiu, Y.-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., et al. (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature.* 402, 404–407. doi: 10.1038/46536

Ratnasingham, S., and Hebert, P. D. (2007). BOLD: The barcode of life data system. *Mol. Ecol. Notes* 1, 355–364. doi: 10.1111/j.1471-8286.2007.01678.x

Sablok, G., Raju, G. V. P., Mudunuri, S. B., Prabha, R., Singh, D. P., Baev, V., et al. (2015). ChloroMitoSSRDB 2.00: more genomes, more repeats, unifying SSRs search patterns and on-the-fly repeat detection database update. *Database.* 84. doi: 10.1093/database/bav084

Satpute, A. D., Chen, C., Gmitter, F. G., Ling, P., Yu, Q., Grosser, M. R., et al. (2015). Cybridization of grapefruit with 'Dancy' mandarin leads to improved fruit characteristics. *J. Am. Soc Hortic. Sci.* 140, 427–435. doi: 10.21273/jashs.140.5.427

Schmitz-Linneweber, C., Lampe, M.-K., Sultan, L. D., and Ostersetzer-Biran, O. (2015). Organellar maturases: A window into the evolution of the spliceosome. *Biochim. Biophys. Acta* 1847, 798–808. doi: 10.1016/j.bbabio.2015.01.009

Sharp, P. (1991). Five easy pieces. *Science.* 254, 663. doi: 10.1126/science.1948046

Shaw, J., Shafer, H. L., Leonard, O. R., Kovach, M. J., Schorr, M., and Morris, A. B. (2014). Chloroplast DNA sequence utility for the lowest phylogenetic and phylogeographic inferences in angiosperms: the tortoise and the hare IV. *Am. J. .Bot.* 101, 1987–2004. doi: 10.3732/ajb.1400398

Shilpa, H. B., and Lohithaswa, H. C. (2021). Discovery of SNPs in important legumes through comparative genome analysis and conversion of SNPs into PCR-based markers. *J. Genet.* 100. doi: 10.1007/s12041-021-01320-3

Sloan, D. B. (2013). One ring to rule them all? genome sequencing provides new insights into the 'master circle' model of plant mitochondrial DNA structure. *New Phytol.* 200, 978–985. doi: 10.1111/nph.12395

Sloan, D. B., Alverson, A. J., Chuckalovcak, J. P., Wu, M., McCauley, D. E., Palmer, J. D., et al. (2012). Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PloS Biol.* 10. doi: 10.1371/journal.pbio.1001241

Sloan, D. B., Oxelman, B., Rautenberg, A., and Taylor, D. R. (2009). Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe sileneae (Caryophyllaceae). *BMC Evol. Biol.* 9, 260. doi: 10.1186/1471-2148-9-260

Soranzo, N., Provan, J., and Powell, W. (1999). An example of microsatellite length variation in the mitochondrial genome of conifers. *Genome* 42, 158–161. doi: 10.1139/g98-111

Sperisen, C., Buchler, U., Gugerli, F., Matyas, G., Geburek, T., and Vendramin, G. G. (2001). Tandem repeats in plant mitochondrial genomes: application to the analysis of population differentiation in the conifer Norway spruce. *Mol. Ecol.* 10, 257–263. doi: 10.1046/j.1365-294X.2001.01180.x

Sugiyama, Y., Watase, Y., Nagase, M., Makita, N., Yagura, M., Hirai, A., et al. (2005). The complete nucleotide sequence and multipart organization of the tobacco mitochondrial genome: comparative analysis of mitochondrial genomes in higher plants. *Mol. Genet. Genom.* 272, 603–615. doi: 10.1007/s00438-004-1075-8

Tonti-Filippini, J., Nevill, P. G., Dixon, K., and Small, I. (2017). What can we do with 1000 plastid genomes? *Plant J.* 90, 808–818. doi: 10.1111/tpj.13491

Tsujimura, M., and Terachi, T. (2018). "Cytoplasmic genome," in *The allium genomes*. Eds. M. Shigyo, A. Khar and M. Abdelrahman (Switzerland: Springer International Publishing), 89–98. doi: 10.1007/978-3-319-95825-5_6

Unseld, M., Marienfeld, J. R., Brandt, P., and Brennicke, A. (1997). The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nat. Genet.* 15, 57–61. doi: 10.1038/ng0197-57

Van Deynze, A., Stoffel, K., Robin, C. R., Kozik, A., Liu, J., van der Knaap, E., et al. (2007). Diversity in conserved genes in tomato. *BMC Genomics* 8, 465. doi: 10.1186/1471-2164-8-465

Vincze, T., Posfai, J., and Roberts, R. J. (2003). NEBcutter: A program to cleave DNA with restriction enzymes. *Nucleic Acids Res.* 31, 368836–368891. doi: 10.1093/nar/gkg526

Wang, Y., Chen, J., Francis, D. M., Shen, H., Wu, T., and Yang, W. (2010). Discovery of intron polymorphisms in cultivated tomato using both tomato and arabidopsis genomic information. *Theor. Appl. Genet.* 121, 1199–1207. doi: 10.1007/s00122-010-1381-y

Wang, X., Wang, J., Li, S., Lu, C., and Sui., N. (2022). An overview of RNA splicing and functioning of splicing factors in land plant chloroplasts. *RNA Biol.* 19, 897–907. doi: 10.1080/15476286.2022.2096801

Wang, X., Zhao, X., Zhu, J., and Wu, W. (2005). Genome-wide investigation of intron length polymorphisms and their potential as molecular markers in rice (*Oryza sativa* l.). *DNA Res.* 12, 417–427. doi: 10.1093/dnares/dsi019

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Wu, G. A., Terol, J., Ibanez, V., López-García, A., Pérez-Román, E., Borredá, C., et al. (2018). Genomics of the origin and evolution of citrus. *Nature.* 554, 311–316. doi: 10.1038/nature25447

Xiang, Q.-P., Wei, R., Shao, Y.-Z., Yang, Z.-Y., Wang, X.-Q., and Zhang, X.-C. (2014). Phylogenetic relationships, possible ancient hybridization, and biogeographic history of abies (Pinaceae) based on data from nuclear, plastid, and mitochondrial genomes. *Mol. Phylogenet. Evol.* 82 Pt A, 1–14. doi: 10.1016/j.ympev.2014.10.008

Xiong, Y., Yu, Q., Yiong, Y., Zhao, J., Lei, X., Liu, L., et al. (2022). The complete mitogenome of *Elymus sibiricus* and insights ito its evolutionary pattern based on simple repeat sequences of seed plant mitogenomes. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.802321

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinf.* 13, 134. doi: 10.1186/1471-2105-13-134

Zimmerly, S., and Semper, C. (2015). Evolution of group II introns. *Mob. DNA* 6. doi: 10.1186/s13100-015-0037-5

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 June 2023
DOI 10.3389/fpls.2023.1205681

OPEN ACCESS

EDITED BY
Yuri Shavrukov,
Flinders University, Australia

REVIEWED BY
Gehendra Bhattarai,
University of Arkansas, United States
Zhansheng Li,
Chinese Academy of Agricultural Sciences, China
Abdur Rahim,
Sher-e-Bangla Agricultural University, Bangladesh

*CORRESPONDENCE
So Young Yi
yisy@kongju.ac.kr
Si-Yong Kang
sykang@kongju.ac.kr
Yong Pyo Lim
yplim@cnu.ac.kr

RECEIVED 14 April 2023
ACCEPTED 24 May 2023
PUBLISHED 13 June 2023

CITATION
Lu L, Choi SR, Lim YP, Kang S-Y and Yi SY (2023) A GBS-based genetic linkage map and quantitative trait loci (QTL) associated with resistance to *Xanthomonas campestris* pv. *campestris* race 1 identified in *Brassica oleracea*. *Front. Plant Sci.* 14:1205681. doi: 10.3389/fpls.2023.1205681

A GBS-based genetic linkage map and quantitative trait loci (QTL) associated with resistance to *Xanthomonas campestris* pv. *campestris* race 1 identified in *Brassica oleracea*

Lu Lu[1], Su Ryun Choi[1], Yong Pyo Lim[2*], Si-Yong Kang[3,4*] and So Young Yi[1,4*]

[1]Institute of Agricultural Science, Chungnam National University, Daejeon, Republic of Korea, [2]Molecular Genetics and Genomics Laboratory, Department of Horticulture, Chungnam National University, Daejeon, Republic of Korea, [3]Department of Horticulture, College of Industrial Sciences, Kongju National University, Yesan, Republic of Korea, [4]Research Center of Crop Breeding for Omics and Artificial Intelligence, Kongju National University, Yesan, Republic of Korea

The production of *Brassica oleracea*, an important vegetable crop, is severely affected by black rot disease caused by the bacterial pathogen *Xanthomonas campestris* pv. *campestris*. Resistance to race 1, the most virulent and widespread race in *B. oleracea*, is under quantitative control; therefore, identifying the genes and genetic markers associated with resistance is crucial for developing resistant cultivars. Quantitative trait locus (QTL) analysis of resistance in the F_2 population developed by crossing the resistant parent BR155 with the susceptible parent SC31 was performed. Sequence GBS approach was used to develop a genetic linkage map. The map contained 7,940 single nucleotide polymorphism markers consisting of nine linkage groups spanning 675.64 cM with an average marker distance of 0.66 cM. The $F_{2:3}$ population (N = 126) was evaluated for resistance to black rot disease in summer (2020), fall (2020), and spring (2021). QTL analysis, using a genetic map and phenotyping data, identified seven QTLs with LOD values between 2.10 and 4.27. The major QTL, *qCaBR1*, was an area of overlap between the two QTLs identified in the 2nd and 3rd trials located at C06. Among the genes located in the major QTL interval, 96 genes had annotation results, and eight were found to respond to biotic stimuli. We compared the expression patterns of eight candidate genes in susceptible (SC31) and resistant (BR155) lines using qRT-PCR and observed their early and transient increases or suppression in response to *Xanthomonas campestris* pv. *campestris* inoculation. These results support the involvement of the eight candidate genes in black rot resistance. The findings of this study will contribute towards marker-assisted selection, additionally the functional analysis of candidate genes may elucidate the molecular mechanisms underlying black rot resistance in *B. oleracea*.

KEYWORDS

Brassica oleracea*, black rot, quantitative trait loci, genotyping by sequencing, linkage map, *Xanthomonas campestris* pv. *campestris

1 Introduction

Brassica oleracea is a plant species that includes several popular vegetables such as broccoli, cauliflower, kale, Brussels sprouts, and cabbage, which are grown and consumed worldwide. Black rot is one of the most prevalent diseases among these crops, caused by the bacterial pathogen *Xanthomonas campestris* pv. *campestris* (*Xcc*). *Xcc* affects *B. oleracea* crops, causing significant economic losses and reducing production performance and quality (Dhar and Singh, 2014). The disease can be transmitted through infected seeds, infested soil, crop residues, and various environmental and mechanical means (Vicente and Holub, 2013). *Xcc* can infect plants at any stage of development and enter through hydathodes, wounds, and other entry points. Symptoms include V-shaped yellow lesions progressing from the leaf margins to the middle vein, darkening of the veins and vascular tissue, premature leaf fall, and stunted growth (Agrios and Dawson, 2004). *Xcc* has 11 physiological races (Vicente et al., 2001; Cruz et al., 2017), with races 1 (R1) and 4 (R4) being the most aggressive and predominant worldwide (Lema et al., 2012; Vicente and Holub, 2013). Control management for black rot is limited but includes copper application, crop rotation, crop debris, cruciferous weed removal, seed treatment, and cultivation of resistant cultivars (Vicente and Holub, 2013). Among all, growing *Xcc*-resistant cultivars can help achieve sustainable and effective disease control. However, according to previous reports, the *R* genes that confer resistance to R1 are present in the B genomes of *Brassica carinata* (BBCC), *Brassica juncea* (AABB), and *Brassica nigra* (BB) (Vicente et al., 2001; Taylor et al., 2002). Currently, no major genes have been identified to be responsible for resistance to black rot in cabbage cultivars. The black rot resistance of most *B. oleracea* lines is considered to be under quantitative control (Taylor et al., 2002; Lema et al., 2012).

The identification of Quantitative Trait Loci (QTL) associated with black rot resistance in cabbage has been a subject of research for several years. Since *Xcc* R1 and R4 are considered the most virulent and widespread races in *B. oleracea*, several studies have been conducted to identify the *R*-genes/QTLs and markers linked to black rot (*Xcc* R1 and R4) resistance in *B. oleracea*. Sharma et al. identified the black rot resistance locus Xca1bc on LG B-7 in Indian mustard (*Brassica carinata*). They also reported that a single dominant gene controls black rot resistance in *B. carinata* (Sharma et al., 2016). Two sequence characterized amplified region (SCAR) markers, ScOPO-04833 and ScPKPS-11635, were identified in close linkage with the black rot resistance locus (*Xca1Bo*) in cauliflower and showed 100% accuracy in differentiating the resistant and susceptible plants of cauliflower breeding lines (Kalia et al., 2017). The identification of QTLs related to resistance to R1 of *Xcc* in cabbage has been a subject of research for several years (Camargo et al., 1995; Doullah et al., 2011; Kifuji et al., 2012; Tonu et al., 2013; Lee et al., 2015; Afrin et al., 2018; Iglesias-Bernabe et al., 2019). QTLs controlling the resistance to *Xcc* R1 have been mapped, and two significant QTLs have been identified in LG2 and LG9 in *B. oleracea* (Camargo et al., 1995; Doullah et al., 2011; Tonu et al., 2013). Kifuji et al. reported a QTL for black rot resistance located on LG C02, which comprises the major QTL in cabbage (Kifuji et al., 2012). Furthermore, researchers have mapped the *Xcc* R1 resistance locus, *Xca1bo*, on chromosome 3 in Indian cauliflowers using bulk segregant analysis, while several random amplified polymorphic DNA (RAPD) markers have been linked to *Xcc* R1 resistance locus (Saha et al., 2014a; Saha et al., 2014b). Lee et al. reported a genetic linkage map where they improved the resolution of a previously developed genetic map, and QTL analysis identified one major (*BRQTL-C1_2*) and three minor QTLs (*BRQTL-C1_1*, *BRQTL-C3*, and *BRQTL-C6*) (Lee et al., 2015). Iglesias-Bernabé et al. measured five traits related to the initial stages of invasion, success of infection, and spread of the pathogen in a BolTBDH mapping population and identified four single-trait QTLs that confirmed the quantitative nature of *Xcc* R1 resistance in linkage groups 1, 6, 8, and 9 (Iglesias-Bernabe et al., 2019).

In a previous study (Lu et al., 2021), we selected an *Xcc*-resistant line, Black rot Resistance 155 (BR155), and a susceptible line, SC31, by comparing symptom development. Using these two cabbage lines, we studied the early defense mechanisms of *B. oleracea* in response to *Xcc* infection and found that BR155 had a relatively strong antioxidant activity. These results suggest that regulating ROS accumulation during early *Xcc*–cabbage interactions may be essential for restricting symptom development. In this study, to identify QTLs for *Xcc* R1 resistance in BR155, we used a reference-based genotyping by sequencing (GBS) approach for single nucleotide polymorphism (SNP) identification and genotyping of a mapping population. The identified SNPs were used to construct linkage maps and to detect loci associated with black rot resistance.

2 Methods

2.1 Plant materials

In a previous study, two inbred cabbage lines (*Brassica oleracea L.* var. *capitate*), SC31 and BR155, showed susceptibility and high resistance to *Xcc* R1, respectively (Lu et al., 2021). In the current study, they were utilized as parents to generate a segregating population (F_1, F_2, and $F_{2:3}$). A total of 126 F_2 individuals were generated and self-pollinated to generate $F_{2:3}$ progenies, and the F_2 and $F_{2:3}$ populations were used for genotyping and phenotypic evaluation, respectively. The parental lines and F_1 and F_2 plant materials examined in this study were obtained from Joeun Seeds Co. (Chungcheongbuk-Do, Korea), and F_2 progenies were self-pollinated to produce seeds of $F_{2:3}$ progenies in a greenhouse facility located at Chungnam National University.

2.2 Phenotypic screening and disease evaluation

Xanthomonas campestris pv. *campestris* KACC 10377 (*Xcc* R1) was used for the inoculation tests in this study, which was obtained from the Korean Agricultural Culture Collection (KACC; Suwon,

Korea). The inoculum and inoculation protocol was conducted as described previously (Lee et al., 2020) with minor modification. Shortly, using an inoculating loop, the bacterial inoculum was streaked over tryptic soy agar (TSA) plates and incubated for 48 hours at 30°C. To prepare the bacterial solution for inoculation, cultivated bacteria were suspended in distilled water and diluted to an optical density (OD) of 0.125 at 600 nm. $F_{2:3}$ seeds were sown and grown on 5×8-cell plug trays in a greenhouse. At 14–17 days after sowing, inoculation tests were performed until the plants had two completely expanded true leaves. The leaves were inoculated by spraying a bacterial suspension until the adaxial and abaxial surfaces of the leaves were sufficiently moistened. Subsequently, the inoculated plants were kept at a temperature of 28°C and high humidity for 48 h. Then, the temperature was adjusted to 25°C and other conditions were kept constant for a further 7 days of incubation, and the disease symptoms in two inoculated leaves of each plant were surveyed. The severity of the black rot symptoms was determined based on the infected leaf area using the following disease indexes: (0) no visible symptoms (immune I), (1) 1–25% infection (resistant, R), (2) 26–50% infection (moderately resistant MR), (3) 51–75% infection (moderately susceptible, MS), (4) 76–99% infection (susceptible, S), (5) 100% infection (highly susceptible, HS) (Peňázová et al., 2018).

2.3 Statistical analysis

SPSS software (v. 26.0, IBM, Armonk, NY, USA) was used for descriptive statistics and correlation analyses of each trial. The mean values of the three trials for each test were used to conduct correlation analyses. The coefficient of variation was calculated as σ/μ, where σ represents the standard deviation and μ represents the average.

2.4 GBS library preparation, sequencing, and SNP calling

To construct the GBS library for sequencing, the genomic DNAs of the parental lines and F_2 population were isolated from their young leaf tissues via the modified CTAB method. The quantity and quality of the DNA were examined using a NanoDrop ND-1000 (Thermo Fisher Scientific Inc., USA) and agarose gel (1.5%) electrophoresis. For GBS analysis, a library was constructed using 128 genomic DNAs belonging to two parent lines and 126 F_2 populations. The library construction was outsourced to SEEDERS sequencing company (Daejeon, Korea; http://www.seeders.co.kr/). Briefly, construction of the GBS library involved the following steps: adaptor annealing, digestion of genomic DNA using *ApeKI* (New England Biolabs, Ipswitch, MA, USA) restriction enzyme, pooling and purification of ligated products, and PCR amplification. The size distribution of the templates was confirmed by analyzing the PCR-enriched fragments on an Agilent Technologies 2100 Bioanalyzer with a DNA 1000 chip. The quality was then assessed through agarose gel electrophoresis. The barcode sequence was used to separate the raw sequences into individual samples, after which the adapter sequence was removed, trimming the sequence quality. We used cutadapt v.1.8.3 (Martin, 2011) for adapter trimming and the Dynamic Trim (phred score ≥ 20) and the LengthSort (short read length ≥ 25 bp) programs of the SolexaQA v.1.13 package for sequence quality trimming (Cox et al., 2010). To ensure accuracy, we used the consensus sequence of SC31 samples obtained through mapping on the corresponding reference genome (*Brassica oleracea*, v.2.1.28; EnsemblPlants, http://plants.ensembl.org/index.html) in resequencing as the reference sequence for analysis. This decision was made due to significant differences between the reference genome and the germplasm of the population being studied. Subsequently, the clean reads of each F2 individual were aligned to the SC31 consensus sequence reference sequence using the Burrows-Wheeler Aligner (BWA) program v.0.6.1-r104 (Li and Durbin, 2009). To create an SNP matrix, we compared the raw SNPs from 126 samples using SEEDERS' in-house script (Kim et al., 2014). Subsequently, the SNPs were classified into homozygous (SNP read depth ≥ 90%), heterozygous (40% ≤ SNP read depth ≤ 60%), and others (homozygous/heterozygous; could not be distinguished by type), followed by SNP filtering (Table S1).

2.5 SNP genotyping and bin construction

Although GBS can rapidly detect thousands of SNPs, not all SNPs detected by GBS can be used to construct genetic maps for genotyping F2 populations. The noise present in sequencing reads can impact the construction of the linkage map. We conducted the chi-square (χ^2) test on all SNPs to assess any potential segregation distortion, and SNPs with a segregation distortion test score of $p < 0.05$, or those with an abnormal base, were removed from the dataset. Additionally, any genotypes with more than 5% deletions were removed, along with corresponding individuals. Finally, we marked specific SNP positions that can be used for calling SNPs in F_2 individuals. The genotype of F_2 individuals was converted to 2 if the SNP was the same as SC31, the genotype of F_2 individuals was converted to 0 if the SNP was the same as BR155, and the genotype of F_2 individuals was converted to 0 if the SNP was the same as F_1. A sliding-window approach was applied for variant calling errors to calculate the ratio of SNP alleles derived from the two parental lines, BR155 and SC31 (Huang et al., 2009). Genotypic data were scanned using a window size of 15 SNPs and a step size of 1. For each individual, the ratio of the SNP alleles from BR155 to SC31 within the window was calculated. Windows with a sum of 15 SNPs were greater than 24, which were considered from SC31, and less than 6, which were considered from BR155, whereas those with varied sums were classified as heterozygous. Adjacent windows with the same genotype were combined into blocks, and recombinant breakpoints were assumed to be at the boundaries of adjacent blocks with different genotypes. Next, a bin map was generated by aligning and comparing the genotypic maps of individual F_2 plants. Consecutive intervals lacking a recombination event within the population were joined into bins that were used as markers. This process was performed using an R script.

2.6 Genetic map construction and QTL mapping

A linkage map was established from the recombination bins that were used as genetic markers via the JoinMap version 5.0 software (https://kyazma.nl/). The Kosambi mapping function was used to convert the recombination frequencies into genetic distances. The disease index for each F_2 individual was calculated as the mean grade of 10–15 $F_{2:3}$ seedlings. QTLs for *Xcc* resistance were evaluated using a composite interval mapping (CIM) analysis with WinQTL cartographer version 2.5 (Zeng, 1994; Rifkin, 2012).

2.7 RNA extraction and gene expression analysis by quantitative real-time PCR

The infected zones of the leaves were collected at 0, 12, 24, and 48 h after inoculation. For each time point, samples from five leaves were combined and considered biological replicates. The leaves were ground into a powder in liquid nitrogen. Total RNA was extracted using the RNeasy® Plant Mini Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions. The extracted RNA was purified using RNeasy® Plant Mini Columns (Qiagen, Hilden, Germany). cDNAs was synthesized using 1 µg of total RNA. A real-time PCR detection system (Bio-Rad, Hercules, CA, USA) with TB Green® Premix Ex Taq ® (TaKaRa Bio) was used to quantify the gene expression. Sequences of the gene-specific primers used for quantitative real-time PCR (qRT-PCR) are presented in Supplementary Table S5. The internal standard used was 18S rRNA. Each experiment was performed at least thrice. The $2^{-\Delta\Delta Ct}$ method was used to quantify the relative transcript level (Livak and Schmittgen, 2001).

3 Results

3.1 Evaluation of resistance to black rot

We observed phenotypes 7 days after the *Xcc* R1 inoculation of parental lines (SC31 and BR155) and 126 $F_{2:3}$ plants in three environments to evaluate black rot resistance (Figure 1). The average disease score from 10 to 15 plants for $F_{2:3}$ plants was considered as the disease score of each $F_{2:3}$ individual (Table S2). Inoculation tests were repeatedly carried out in the summer and fall of 2020, and in the spring of 2021 under the same conditions (Figure 2). The frequency distribution of the black rot disease index of the $F_{2:3}$ population under the two different environmental

FIGURE 1
Representative black rot disease symptoms on leaves of cabbage after spraying with *Xcc* R1 suspension (OD_{600} = 0.125). The severity of the black rot symptoms was recorded based on infected leaf area, with the following disease indices: (0) no visible symptoms, (1) 1–25% infection (resistant, R), (2) 26–50% infection (moderately resistant MR), (3) 51–75% infection (moderately susceptible, MS), (4) 76–99% infection (susceptible, S), (5) 100% infection (highly susceptible, HS).

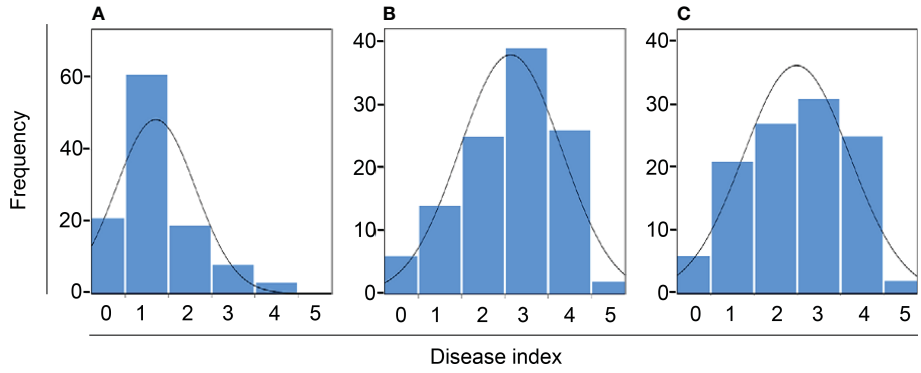


FIGURE 2
Frequency distribution of black rot disease index of $F_{2:3}$. **(A–C)** represent the inoculation test result from summer 2020, fall 2020, and spring 2021, respectively. The curve indicates normal distribution.

conditions appeared to be approximately normally distributed in separate experiments, indicating that the black rot resistance phenotype was governed by multiple genes (Figure 2). The correlation coefficients among the three trials were calculated (Table 1). As a result of the correlation analysis between the three trials using 128 cabbage samples (parents and $F_{2:3}$), the second and third trials had a positive correlation of approximately 0.66 at the level of $\alpha = 0.001$ (***). On the other hand, the first and third trails had a negative correlation of roughly -0.20 (Table 1).

3.2 Whole-genome resequencing of two cabbage parental lines and SNP detection by GBS

The whole-genome sequencing data contained 68 and 67 million raw reads for BR155 and SC31, respectively (Table S3). According to Parkin et al., the genome sequence for *B. oleracea* is 488.6 Mb, consisting of 446.9 Mb of nine pseudo-chromosomes and 41.2 Mb of unanchored scaffolds, accounting for approximately 75% of the estimated genome size of 648 Mb (Parkin et al., 2014). Our new sequencing data was approximately 24 times the genome size for both parent lines. We successfully mapped each set of paired reads onto the nine pseudo-chromosomes of the reference genome sequence. Of the raw reads obtained, 70.74% and 74.67% from BR155 and SC31, respectively, were successfully aligned to the reference genome, which resulted in a mapped region of 72.44% and 78.39% for BR155 and SC31, respectively (Table S3). The total number of SNPs and average SNP densities varied between the two parental lines. High-quality SNPs were identified in both BR155 and SC31, with approximately 1.02 million and 0.26 million SNPs, respectively. These SNPs were merged and used to detect SNPs between the two parental lines (Table 2).

For the genome-wide detection of SNPs in cabbage using GBS, the restriction enzyme *ApeKI* was used to digest genomic DNA and construct GBS libraries of the F_2 plants and parents of the intraspecific mapping population (BR155 and SC31). Sequencing was performed on an Illumina high-throughput sequencing platform (Illumina HiSeq X sequencer), and a total of 1,501,825,142 raw sequence reads corresponding to 226.78 GB of sequence length were generated. The raw data contained an average of 97.11% of demultiplexed reads, with the overall GC content of the sequences being approximately 47.66%, And the Q30 score was approximately 91.29%. Raw SNPs were detected by sequence pre-processing and alignment to the SC31 consensus sequence, and a matrix containing 304,184 SNPs was obtained. The SC31 sequence was aligned to the *B. oleracea* (TO1000) sequence to determine the physical position of each SNP. Based on the filtering process, 27,403 polymorphic SNPs were identified in the cabbage F2 population (Table 2). SNPs were distributed across all nine *B. oleracea* chromosomes, as illustrated in Figure 3 and Table 3.

TABLE 1 Phenotypic correlation for black rot disease index over summer 2020, fall 2020, and spring 2021.

Trial	1st (Summer 2020)	2nd (Fall 2020)	3rd (Spring 2021)
1st (Summer 2020)	1		
2nd (Fall 2020)	-0.13	1	
3rd (Spring 2021)	-0.2	0.66***	1

***Significant at the 0.001 probability level.

3.3 Development of the linkage map for *Brassica oleracea*

After filtering the SNPs according to the genotyping criteria, 7,940 high-quality SNPs were identified between the two parents to generate bin markers for the F_2 population (Table 3). A modified sliding window approach was adopted to determine the recombinant breakpoints for the F_2 individuals. The adjacent bins of the same genotype were merged into identical bins. A high-density genetic linkage map was constructed using 1,020 recombination bins (Figure 4, Table 3). The total genetic distance of the linkage map was 675.64 cM with an average distance of 0.66 cM between adjacent bins. Linkage Group 3 (C03) contained the most bins (188), followed by Group C09 (125). Group 3 also comprised the longest linkage group, which spanned 114.32 cM and 95.29 cM with an average 0.61 cM and 0.76 cM marker intervals, respectively. The shortest linkage group was located at C06, which was 49.02 cM in length and harbored 87 bin markers with an average marker interval of 0.56 cM.

3.4 QTL analysis

Composite interval mapping was conducted using the developed cabbage map to detect black rot resistance QTL. QTL analysis was performed for each trial. QTLs were detected based on LOD scores higher than the threshold (2.0). As a result, a total of seven QTLs were detected in this study, the LOD values between 2.10 and 4.27 (Figure 4, Table 4; Figure S1). In the first test, performed in the summer of 2020, there were two significant QTL regions: *qCaBR-C2-1* on chromosome C02 and *qCaBR-C5* on chromosome C05 (Figure 4, Table 4; Figure S1). Among these, *qCaBR-C5* showed the highest LOD score, 4.27 (Figure 4, Table 4; Figure S1). In fall 2020, the second test identified three QTLs, *qCaBR-C2-2*, *qCaBR-C6-1*, and *qCaBR-C7* on chromosomes C02, C06, and C07, respectively. Two QTLs were detected in spring 2021: *qCaBR-C2-3* and *qCaBR-C6-2*.

3.5 Prediction of candidate genes for black rot resistance in *B. oleracea*

Based on the above analysis, among all seven QTLs (Figure 4, Table 5), we found overlapping regions between the second and third trials located on chromosome C06 which were designated as the major QTL loci. The major QTL identified on chromosome 6 was *qCaBR1* (Cabbage Black rot Resistance-1) (Figure 5). According to the available *B. oleracea* (TO1000) genome sequence (http://plants.ensembl.org/

TABLE 2 Distribution of SNPs on the *B. oleracea* genome.

Chr.	Chr. Length (bp)	No. genes[a]	Raw SNPs	Filtered SNP	Selected SNPs[b]
C01	43,764,888	5,401	44,427	1,940	528
C02	52,886,895	5,843	112,580	3,811	1,075
C03	64,984,695	8,490	89,494	3,618	1,040
C04	53,719,093	6,426	74,076	2,718	868
C05	46,902,585	5,851	71,841	2,931	828
C06	39,822,476	4,762	70,502	2,636	784
C07	48,366,697	5,752	94,037	3,645	1,030
C08	41,758,685	5,599	74,419	2,980	798
C09	54,679,868	6,642	76,718	3,124	989
Total	446,885,882	54,766	708,094	27,403	7,940

[a]Data source: EnsemblPlants (http://plants.ensembl.org/index.html).
[b]We removed F_2 plants with > 25% missing and SNPs with missing in filtered SNPs.

index.html), *qCaBR1* was detected on C06:29,853,043-34,373,426 (4.52 Mb) (Figure 5, Table 5) and included 591 genes. Of the 591 *B. oleracea* genes, 96 (Table S4) had putative gene annotation data (Arabidopsis orthologs), for which we could categorize the functional groups (https://www.arabidopsis.org). According to the annotation information, eight out of 96 genes responded to biotic stimuli and were related to defense responses against other organisms (bacteria, fungi, and oomycetes) (Table 6). We compared the expression patterns of eight candidate genes (Figure 6) that may be related to black rot resistance after *Xcc* R1 inoculation (12, 24, and 48 h) in the two cabbage lines, BR155 and SC31, to assess their defense-related responses. In BR155, two defense-related genes, *PR1* and *SOD*, were rapidly and strongly induced by *Xcc* inoculation, and thus can be used as early defense response markers (Lu et al., 2021). In this study, we also observed that the *Xcc*-induced expression levels of *PR1* and *SOD* was more than two times higher than that in BR155 plants when compared to SC31 plants. A high level of relative expression was observed for four genes (Bo6g098480, Bo6g099850, Bo6g101010, and Bo6g106440) at all time points in BR155; these results were similar to those of *PR1* and *SOD*. However, the expression patterns of Bo6g095580 and Bo6g101310 were opposite those of *PR1* and *SOD*. As shown in Figure 6, the *Xcc*-induced expression levels of Bo6g095580, Bo6g101310, and Bo6g101210 were higher in the susceptible line SC31 than in BR155 (resistant line). Interestingly, the gene expression pattern of Bo6g108870 significantly increased only 24 h post-inoculation in BR155, a resistant parental line (Figure 6). These eight genes showed differential expression patterns between the BR155 and SC31 plants in response to *Xcc* inoculation. Thus, the qRT-PCR results indicate that all eight genes selected from the major QTL interval may be involved in the black rot resistance of BR155.

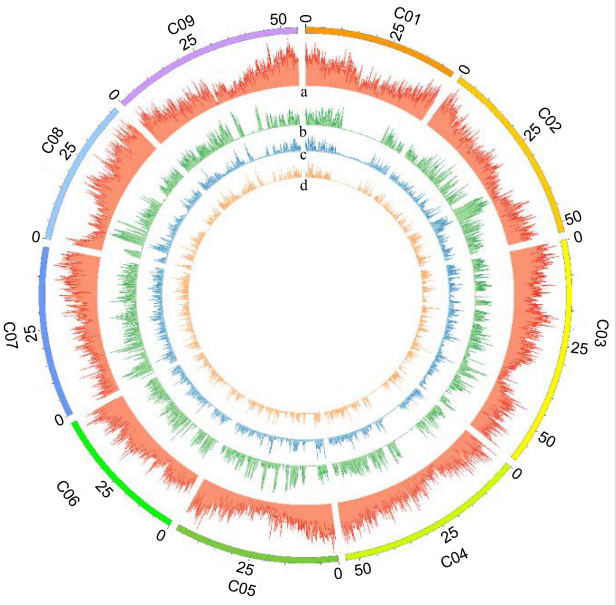


FIGURE 3
Distribution of identified SNPs on nine *B. oleracea* chromosomes. Distribution of SNPs detected on each *B. oleracea* chromosome (100 kb window size) was visualized using Circos diagram-Perl script. The outermost circle represents nine *B. oleracea* chromosomes designated as "C01-C09". Panels as seen from outside to inside: (a) gene density, (b) raw SNPs, (c) filtered SNPs, and (d) selected SNPs for mapping on each chromosome.

3 Discussion

The resistant line "BR155" and susceptible line "SC31" used in this study were selected from 94 *B. oleracea* lines by comparing the lesion areas after pathogenicity assays using the scissor-clipping method (Lu et al., 2021). SC31 was one of 23 lines with a symptom area of 90% or more, and BR155 was the most resistant cabbage line with a lesion area of <10%. Previous studies have indicated that BR155 may carry a highly effective resistance gene or locus. We compared the two cabbage lines for the *Xcc*-induced expression pattern of 13 defense-related genes. Among them, the *Xcc*-induced expression levels of *PR1* and antioxidant-related genes (*SOD*, *POD*, *APX*, *Trx H*, and *CHI*) in BR155 were over twice as high as those in SC31. Nitroblue tetrazolium (NBT) and diaminobenzidine

TABLE 3 Distribution of bin markers on the cabbage genetic map.

Linkage group (Ch)	Bin No.	Genetic length (cM)	Average bins interval (cM)
C01	83	59.43	0.72
C02	97	64.67	0.67
C03	188	114.32	0.61
C04	110	87.90	0.80
C05	115	79.90	0.69
C06	87	49.02	0.56
C07	119	64.56	0.54
C08	96	60.55	0.63
C09	125	95.29	0.76
Total	1,020	675.64	0.66

tetrahydrochloride (DAB) staining analysis showed that BR155 accumulated less *Xcc*-induced reactive oxygen species (ROS) than did SC31. Furthermore, 2, 2-diphenyl-1-picrylhydrazyl (DPPH) radical scavenging assays showed that BR155 had higher antioxidant activity than SC31 (Lu et al., 2021). Identifying the resistance locus of BR155 will be crucial for understanding the mechanism of black rot disease resistance. Therefore, in the present study, a GBS-based genetic linkage map was developed and QTL linked to resistance against *Xcc* R1 in cabbage was identified.

Genotyping-by-sequencing (GBS) technology allows for efficient and cost-effective genotyping of large numbers of markers across the genome. Here, we applied high-throughput GBS technology with the type-II restriction endonuclease *ApeK*I (Elshire et al., 2011) to the F_2 group of cabbage, enabling the simultaneous identification of sufficient polymorphic SNPs and genotyping. This allowed us to create a linkage map with reasonably high density without the need to check or apply existing markers (Figure 4). A modified sliding window approach was adopted to determine the recombinant breakpoints for the F_2 individuals. The sliding-window approach involves calculating the recombination frequency between adjacent markers within each window. Recombination frequency measures how often genetic recombination occurs between two markers during meiosis, and can be used to estimate the physical distance between them on the chromosome (Tang et al., 2009; Beissinger et al., 2015). The GBS analysis performed by Parkin et al. identified 826 bins in *B. oleracea* (Parkin et al., 2014), which was fewer than that identified in this study (1020 bins). Our higher bin numbers are probably due to the difference in the genetic diversity of the parental lines and the number of segregating progenies used for GBS analysis. We used BR155 and SC31 as parental lines, which have variable genetic diversity, whereas Parkin et al. analyzed the population of double haploid (DH) kale-like and DH broccoli lines. We also analyzed 126 F_2 plants for mapping. In this study, we identified 708,094 SNPs between BR155 and SC31 using 24x genome coverage whole-genome resequencing (Tables 2; S3). Here, 27,403 GBS-based SNPs were detected between the parental lines, which was 7-fold fewer than those detected by resequencing. After filtering the SNPs according to the genotyping criteria, 7,940 high-quality SNPs were identified between the two parents to generate bin markers for the F_2 population (Table 2). The total genetic distance of the linkage map was 675.64 cM, with an average length of 0.66 cM between adjacent bins (Table 3).

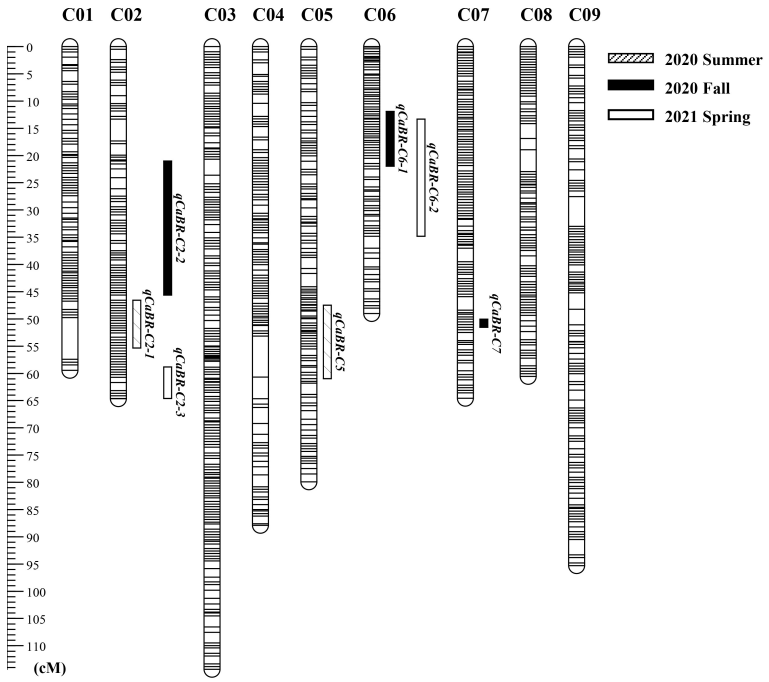


FIGURE 4
The genetic linkage map of cabbage constructed using the F_2 lines derived from the parental lines SC31 × BR155. Detected QTLs were marked by rectangular bars with different fill in the right side of the linkage groups followed by QTL name.

10.3389/fpls.2023.1205681

TABLE 4 Details of resistance QTLs related *Xcc* R1.

Inoculation test	QTL name	Linkage group	Number of bin	Position (cM)	Physical position (bp)	LOD[a]	Add[b]	Dominance effect	R^2 (%)[c]
(1st) 2020 Summer	*qCaBR-C2-1*	C02	19	46.57 - 54.86	16,499,782-49,025,692	2.55	0.37	-0.25	8.2
	qCaBR-C5	C05	27	47.47 - 60.98	12,846,418-42,204,163	4.27	-0.46	-0.45	13.6
(2nd) 2020 Fall	*qCaBR-C2-2*	C02	38	21.02 - 45.62	3,513,666-16,499,781	3.42	0.56	0.19	12.2
	qCaBR-C6-1	C06	23	11.93 - 21.96	27,081,550-34,373,426	2.18	-0.27	-0.51	7.28
	qCaBR-C7	C07	4	50.05 - 51.55	44,486,293-45,629,195	2.1	0.4	0.33	6.75
(3rd) 2021 Spring	*qCaBR-C2-3*	C02	6	59.30 - 63.15	50,467,819-52,028,484	2.39	-0.22	0.65	8.03
	qCaBR-C6-2	C06	39	13.32 - 34.84	29,853,043-37,960,957	2.76	-0.56	-0.38	9.9

[a]Logarithm of odds ratio at the position of the peak.
[b]Additive effect of QTL.
[c]Percent of phenotypic variance explained by the QTL.

We evaluated black rot resistance by observing phenotypes seven days after inoculating parental lines (SC31 and BR155) and 126 $F_{2:3}$ plants with *Xcc* R1. The inoculation tests were performed three times throughout the summer and fall of 2020, as well as in the spring of 2021, all under identical conditions (Figure 2). As shown in Table 1, analysis of the correlation between three trials consisting of 128 cabbage samples, including parents and $F_{2:3}$, revealed that the second and third trials had a strong positive correlation of approximately 0.66 at a significance level of $\alpha = 0.001$ (***). In contrast, the first and third trials displayed a negative correlation of approximately -0.20. It seems possible that such an unexpected result in the first trials' outcome was affected by the high temperature during summer. In 1972, Staub and Williams analyzed the impact of temperature on the black rot resistance of cabbage by exposing resistant and susceptible cabbage varieties inoculated with *Xcc* to various temperatures and analyzing the severity of black rot. Their results showed that although the traits of the resistant cabbage were evident at 20-24°C, the resistant cabbage was just as susceptible to the disease as the susceptible ones at 28°C (Staub, 1972). In Korea, the temperature inside a greenhouse can exceed 30°C during the summer.

Xcc exhibits high genetic diversity, and 11 races have been discovered worldwide. Among these races, R1 and R4 are the most prevalent and highly virulent among many commercial cultivars (Vicente et al., 2001; Cruz et al., 2017). Only a few resistant resources have been identified recently, considerably challenging the breeding of resistant cabbage cultivars. Several studies have identified R-genes/QTLs and markers associated with *Xcc* R1 and *Xcc* R4 resistance in *B. oleracea* (Camargo et al., 1995; Kifuji et al., 2012; Tonu et al., 2013; Saha et al., 2014b; Lee et al., 2015; Iglesias-Bernabe et al., 2019). In total, more than 15 QTLs were identified on *the B. oleracea* chromosomes, indicating that resistance to black rot is highly complicated. We identified seven QTLs related to the resistance to R1 of *Xcc* on C02, C05, C06, and C07. Since resistance was quantitative and under polygenic control, we confirmed the results of other studies. The positions of our black rot resistance QTLs coincided with those previously reported (Kifuji et al., 2012; Tonu et al., 2013; Saha et al., 2014b; Iglesias-Bernabe et al., 2019), except for QTL *qCaBR*-C7, which may represent a novel variation. Among the seven QTLs identified in this study, *qCaBR-C6-1* and *qCaBR-C6-2* were detected repeatedly in the two independent inoculation tests, had high LOD values, and accounted for a high percentage of the variation in all trials. We designated the overlapping part of these two QTLs on C06 as *qCaBR1* QTL. In addition, it was a strong candidate as a major QTL for black rot resistance. Regarding physical location, the major QTL, *qCaBR1*, found in our work is likely related to *BRQTL-C6* (Lee et al., 2015) and *Xcc*6.1 (Iglesias-Bernabe et al., 2019). Afrin selected five markers capable of distinguishing the resistant lines from the susceptible ones of cabbage consistently (Afrin et al., 2018). The SSR marker OI10G06 is one of these five markers. Interestingly, OI10G06 is located on chromosome C06 (C6:29898028-29898121) and is in the major QTL *qCaBR1* (C06:29,853,043–34,373,426) (Table 5). In the case of *BRQTL-C6*, the exact *Xcc* race used in the study is yet to be classified, and OI10G06 was able to separate resistant and susceptible lines but did not perfectly match the phenotypic data. However, these repeated reports related to *Xcc* resistance QTL strongly, supporting the idea that the QTL *qCaBR1* is involved in resistance to *Xcc* R1.

The information on the QTLs identified in this study will assist in the understanding of the molecular mechanisms of disease response in

TABLE 5 The physical position of major QTL related to black rot resistance in *B. oleracea*.

QTL name	QTL position in linkage group (cM)	Physical position (bp)	A	B	C
qCaBR1	C06:13.32 - 21.96	C06: 29,853,043 - 34,373,426	591	96	8

A, Number of predicted genes gene in QTL region.
B, Number of genes with functional annotation in QTL region.
C, Number of genes associated with response to biotic stimulus in QTL region.

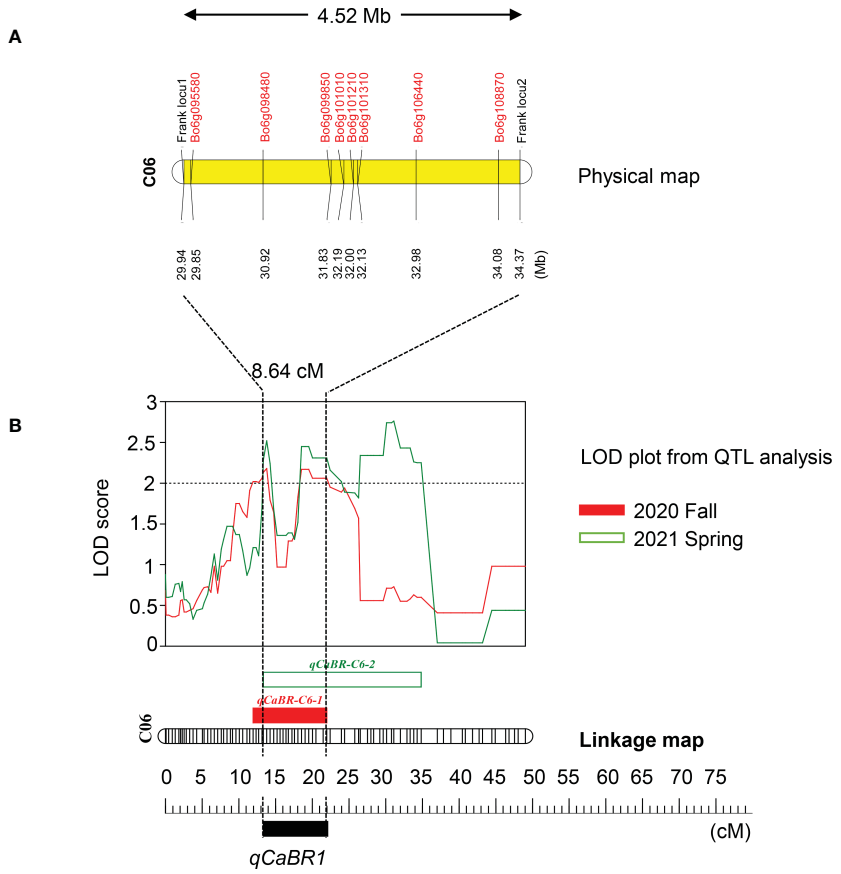


FIGURE 5
Map of the corresponding qCaBR1 region of the B. oleracea genome. **(A)** The corresponding physical map showed the location of the major QTL region in the B. oleracea reference genome and the included reference genes associated with biotic stresses. **(B)** The major QTL was named qCaBR1 as the overlapping section, detected on linkage group six in fall 2020 and spring 2021, respectively. The genetic position of QTLs were indicated in centimorgans (cM) and marked by rectangular bars with a different color in the left side of the linkage group followed by the QTL name. The change curves of LOD values obtained from the results of the three inoculation experiments were shown in red, green, and yellow for fall 2020, spring 2021, and the overlapping section, respectively.

B. oleracea under *Xcc* stress. According to the available *B. oleracea* genome sequence (http://plants.ensembl.org/), the *qCaBR1* locus was delimited to a 4.53-Mb genomic region, which included 96 functionally annotated Arabidopsis orthologs. We identified candidate genes within this chromosomal region with the FGENESH online program (http://linux1.softberry.com/berry.phtml), and the NCBI BLASTP algorithm (http://blast.ncbi.nlm.nih.g/blast) (Table S4). We also identified gene ontology terms using the "Go Term Enrichment tool" on the Tair home page (https://www.arabidopsis.org/tools/go_term_enrichment.jsp). According to this analysis, eight of the 96 genes were biotic-stimulus-responsive (Table 6). These genes are involved in the resistance responses to bacterial or fungal diseases (Mobley et al., 1999; Clay et al., 2009; Li et al., 2019; Depuydt and Vandepoele, 2021) and microbe-associated molecular patterns (Park et al., 2015; Didelon et al., 2020). Since these genes were also speculated to play a role in defense responses against invading pathogens in cabbage, we designated them as candidate genes associated with black rot resistance in the BR155 line. To further assess their roles in black rot resistance, we performed an expression analysis of eight candidate genes (Figure 6). As expected, most of the candidate genes tested showed more robust expression than in SC31 from BR155 cells, a resistant line. However, in the case of Bo6g095580 and Bo6g101310, their expression was more strongly induced in SC31, a susceptible line, in response to *Xcc* inoculation. In the case of Bo6g101210, the *Xcc*-induced expression level was initially high in the susceptible lines; however, after 48 h, it increased in the resistant lines (Figure 6). Bo6g098480, Bo6g101310, and Bo6g108870 showed high homology to AT1G67880, AT1G66830, and AT1G69450, respectively. Depuydt and Vandepoele (2021) inferred the functions of several unknown Arabidopsis genes through omics-supported functional annotation analysis and classified AT1G67880, AT1G66830, and AT1G69450 as having functions related to plant disease resistance (Depuydt and Vandepoele, 2021). Bo6g099850 showed high homology with the Arabidopsis ethylene receptor 1 (AT1G66340). AtETR1, an ET receptor, is required for ET perception (Schaller and Bleecker, 1995) and for the microbe-associated molecular pattern (MAMP)-triggered immune response (MTI). A previous study found that etr1-1 and etr1-3 mutants of the ET signaling pathway were impaired in the Flg22-induced callose response, an MTI (Clay et al., 2009). Bo6g101010 showed a high homology of AT1G66480 to Arabidopsis. AT1G66480 is a protein with pathogen and abiotic stress responses, cadmium tolerance, and disordered region-containing (PADRE) domains.

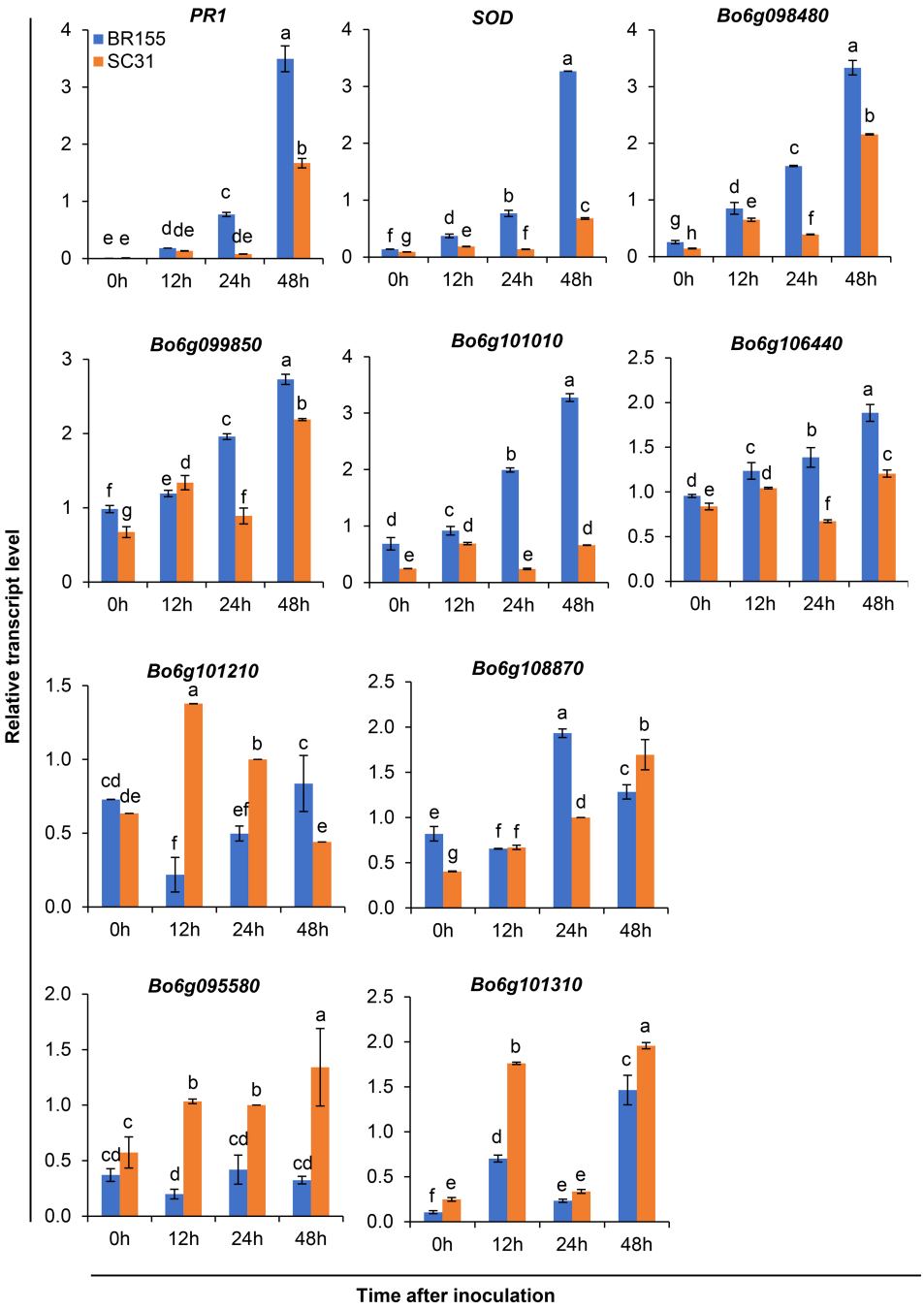


FIGURE 6
Relative transcript levels of defense-related genes (PR1 and SOD) and eight candidate genes in leaves of resistant (BR155) and susceptible (SC31) cabbage lines. Gene expression level was determined by qRT-PCR at the indicated time after Xcc R1 inoculation and normalized to transcript levels of the 18S rRNA gene. Error bars represent standard deviations of three replicates. Similar results were obtained in at least two independent experiments. Different letters indicate significant differences among samples (a = 0.05, one-way ANOVA and Duncan's multiple range test).

Moreover, the latter is in the top 10% of most induced genes after infection with the fungal pathogen *Sclerotinia sclerotiorum* (Didelon et al., 2020). Bo6g106440 shows high homology with Arabidopsis BPL2 (AT1G67950). Arabidopsis accelerated cell death11 (ACD11) encodes a sphingosine transfer protein, and knockout of ACD11 activates PCD and defense responses (Brodersen et al., 2002). BPA1 (the binding partner of ACD11) and its close homologs (BPLs, BPA1-Like proteins) are novel regulators controlling the ROS-mediated defense response and are targeted and manipulated by a virulence effector of Phytophthora, RxLR207. RxLR207 promotes pathogen infection by binding and degrading BPA1 and BPLs (Li et al., 2019). Bo6g095580 had high homology with Arabidopsis Chorismate mutase 3 (AtCM3; AT1G69370). CMs are enzymes that catalyze the conversion of chorismate, a key intermediate in the shikimate pathway, to prephenate, a precursor of the aromatic amino acids phenylalanine and tyrosine (Eberhard et al., 1996). In Arabidopsis, two plastid-localized CMs are allosterically regulated (AtCM1 and AtCM3) and one cytosolic isoform (AtCM2)

TABLE 6 A list of genes associated with response to biotic stimulus located in the major QTL interval.

B. oleracea_ID	Arabidopsis_ID	Description	Response to biotic stimulus	References
Bo6g095580	AT1G69370	Chorismate mutase 3	defense response to bacterium	Mobley et al., 1999
Bo6g098480	AT1G67880	ß-1,4-N-acetylglucosaminyltransferase family protein	defense response to bacterium, fungus	Depuydt and Vandepoele, 2021
Bo6g099850	AT1G66340	Ethylene receptor 1	defense response to bacterium	Clay et al., 2009
Bo6g101010	AT1G66480	Plastid movement impaired 2	response to fungus	Didelon et al., 2020
Bo6g101210	AT1G66730	DNA ligase 6	response to molecule of bacterial origin	Park et al., 2015
Bo6g101310	AT1G66830	Leucine-rich repeat protein kinase family protein	defense response to bacterium, fungus	Depuydt and Vandepoele, 2021
Bo6g106440	AT1G67950	RNA-binding family protein	defense response to fungus	Li et al., 2019
Bo6g108870	AT1G69450	Early-responsive to dehydration stress protein	defense response to other organism	Depuydt and Vandepoele, 2021

is unregulated (Eberhard et al., 1996; Mobley et al., 1999). AtCM3-like isoforms are found only in the Brassicaceae family, suggesting that AtCM3-like isoforms may play a role in specialized metabolite production and stress responses in Brassicaceae (Westfall et al., 2014). For example, indole glucosinolates (IGs) are plant secondary metabolites derived from the amino acid tryptophan found in the family Brassicaceae, and IG synthesis requires indole- and sulfur-containing amino acids and activation of AtCM3 (Grubb and Abel, 2006). Bo6g101210 shares high homology with AT1G66730 in Arabidopsis. AT1G66730 encodes a novel plant-specific DNA ligase, DNA LIGASE VI, which is involved in the response to molecules of bacterial origin (Park et al., 2015). Expression analysis of these eight candidate genes in major QTL intervals and their functional characterization may provide additional molecular information regarding the role of this genomic region in controlling *Xcc* R1 resistance in *B. oleracea*.

In this study, we identified SNPs in the *B. oleracea* genome using a GBS approach. Using the identified SNPs, a linkage map of BR155 and SC31 was constructed. In addition, we mapped one major QTL and seven minor QTLs for *Xcc* R1 resistance. The information generated on QTLs is useful for fine mapping and future MAS of traits. In addition, these results can be applied to the development *Xcc* R1-resistant genotypes and the molecular dissection of *Xcc* R1 resistance in *B. oleracea*.

Data availability statement

The datasets presented in this study can be found in online repositories. The name of the repository and accession number can be found below: NCBI; PRJNA974155.

Author contributions

SY: Data curation. SY: Funding acquisition. LL and SY: Investigation. SC, LL and SY: Methodology. SY: Project administration. YL and S-YK: Resources. LL and SY: Visualization. LL, SY, and S-YK: Manuscript writing. All authors contributed to the article and approved the submitted version.

Funding

This study was supported by the Ministry of Education of the Republic of Korea and National Research Foundation of Korea (grant number NRF-2018K1A3A7A03089858).

Acknowledgments

The authors thank the European Union Horizon 2020 Research and Innovation program for allowing us to use *B. oleracea* germplasm, such as the BRESOV core collection, for this study under grant agreement No 774244 (Breeding for Resilient, Efficient, and Sustainable Organic Vegetable Production; BRESOV).

Conflict of interest

The authors declare that this study was conducted in the absence of any commercial or financial relationships that could be construed as 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.2023.1205681/full#supplementary-material

References

Afrin, K. S., Rahim, M. A., Park, J.-I., Natarajan, S., Rubel, M. H., Kim, H.-T., et al. (2018). Screening of cabbage (Brassica oleracea l.) germplasm for resistance to black rot. *Plant Breed. Biotechnol.* 6 (1), 30–43. doi: 10.9787/pbb.2018.6.1.30

Agrios, G. N., and Dawson, B. (2004). *Plant pathology* (Amsterdam; London: Elsevier Academic).

Beissinger, T. M., Rosa, G. J., Kaeppler, S. M., Gianola, D., and de Leon, N. (2015). Defining window-boundaries for genomic analyses using smoothing spline techniques. *Genet. Sel. Evol.* 47 (1), 30. doi: 10.1186/s12711-015-0105-9

Brodersen, P., Petersen, M., Pike, H. M., Olszak, B., Skov, S., Odum, N., et al. (2002). Knockout of arabidopsis accelerated-cell-death11 encoding a sphingosine transfer protein causes activation of programmed cell death and defense. *Genes Dev.* 16 (4), 490–502. doi: 10.1101/gad.218202

Camargo, L. E. A., Williams, P. H., and Osborn, T. C. (1995). Mapping of quantitative trait loci controlling resistance of brassica oleracea to xanthomonas campestris pv. campestris in the field and greenhouse. *Phytopathol* 85, 1296–1300.

Clay, N. K., Adio, A. M., Denoux, C., Jander, G., and Ausubel, F. M. (2009). Glucosinolate metabolites required for an arabidopsis innate immune response. *Science* 323 (5910), 95–101. doi: 10.1126/science.1164627

Cox, M. P., Peterson, D. A., and Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of illumina second-generation sequencing data. *BMC Bioinf.* 11 (1), 1–6. doi: 10.1186/1471-2105-11-485

Cruz, J., Tenreiro, R., and Cruz, L. (2017). Assessment of diversity of xanthomonas campestris pathovars affecting cruciferous plants in portugal and disclosure of two novel x-campestris pv. Campestris races. *J. Plant Pathol.* 99 (2), 403–414. Available at: https://www.jstor.org/stable/44686785.

Depuydt, T., and Vandepoele, K. (2021). Multi-omics network-based functional annotation of unknown arabidopsis genes. *Plant J.* 108 (4), 1193–1212. doi: 10.1111/tpj.15507

Dhar, S., and Singh, D. (2014). Performance of cauliflower genotypes for yield and resistance against black rot (Xanthomonas campestris pv. campestris). *Indian J. Hortic.* 71 (2), 197–201.

Didelon, M., Khafif, M., Godiard, L., Barbacci, A., and Raffaele, S. (2020). Patterns of sequence and expression diversification associate members of the PADRE gene family with response to fungal pathogens. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00491

Doullah, M. A. U., Mohsin, G. M., Ishikawa, K., Hori, H., and Okazaki, K. (2011). Construction of a linkage map and QTL analysis for black rot resistance in brassica oleracea l. *Int. J. Natural Sci.* 1 (1), 1–6. doi: 10.3329/ijns.v1i1.8591

Eberhard, J., Ehrler, T. T., Epple, P., Felix, G., Raesecke, H. R., Amrhein, N., et al. (1996). Cytosolic and plastidic chorismate mutase isozymes from arabidopsis thaliana: molecular characterization and enzymatic properties. *Plant J.* 10 (5), 815–821. doi: 10.1046/j.1365-313x.1996.10050815.x

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Grubb, C. D., and Abel, S. (2006). Glucosinolate metabolism and its control. *Trends Plant Sci.* 11 (2), 89–100. doi: 10.1016/j.tplants.2005.12.006

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., et al. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19 (6), 1068–1076. doi: 10.1101/gr.089516.108

Iglesias-Bernabe, L., Madloo, P., Rodriguez, V. M., Francisco, M., and Soengas, P. (2019). Dissecting quantitative resistance to xanthomonas campestris pv. campestris in leaves of brassica oleracea by QTL analysis. *Sci. Rep.* 9 (1), 2015. doi: 10.1038/s41598-019-38527-5

Kalia, P., Saha, P., and Ray, S. (2017). Development of RAPD and ISSR derived SCAR markers linked to Xca1Bo gene conferring resistance to black rot disease in cauliflower (Brassica oleracea var. botrytis l.). *Euphytica* 213 (10), 232. doi: 10.1007/s10681-017-2025-y

Kifuji, Y., Hanzawa, H., Terasawa, Y., Ashutosh,, and Nishio, T. (2012). QTL analysis of black rot resistance in cabbage using newly developed EST-SNP markers. *Euphytica* 190 (2), 289–295. doi: 10.1007/s10681-012-0847-1

Kim, J.-E., Oh, S.-K., Lee, J.-H., Lee, B.-M., and Jo, S.-H. (2014). Genome-wide SNP calling using next generation sequencing data in tomato. *Mol. Cells* 37 (1), 36. doi: 10.14348/molcells.2014.2241

Lee, S. M., Choi, Y. H., Kim, H. T., and Choi, G. J. (2020). Development of an efficient screening method for resistance of Chinese cabbage cultivars to black rot disease caused by xanthomonas campestris pv. campestris. *Hortic. Sci. Technol.* 38 (4), 547–558. doi: 10.7235/Hort.20200051

Lee, J., Izzah, N. K., Jayakodi, M., Perumal, S., Joh, H. J., Lee, H. J., et al. (2015). Genome-wide SNP identification and QTL mapping for black rot resistance in cabbage. *BMC Plant Biol.* 15, 32. doi: 10.1186/s12870-015-0424-6

Lema, M., Cartea, M. E., Sotelo, T., Velasco, P., and Soengas, P. (2012). Discrimination of xanthomonas campestris pv. campestris races among strains from northwestern Spain by brassica spp. genotypes and rep-PCR. *Eur. J. Plant Pathol.* 133 (1), 159–169. doi: 10.1007/s10658-011-9929-5

Li, Q., Ai, G., Shen, D. Y., Zou, F., Wang, J., Bai, T., et al. (2019). A phytophthora capsici effector targets ACD11 binding partners that regulate ROS-mediated defense response in arabidopsis. *Mol. Plant* 12 (4), 565–581. doi: 10.1016/j.molp.2019.01.018

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

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2– ΔΔCT method. *methods* 25 (4), 402–408. doi: 10.1006/meth.2001.1262

Lu, L., Monakhos, S. G., Lim, Y. P., and Yi, S. Y. (2021). Early defense mechanisms of brassica oleracea in response to attack by xanthomonas campestris pv. campestris. *Plants-Basel* 10 (12), 2705. doi: 10.3390/plants10122705

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17 (1), 10–12. doi: 10.14806/ej.17.1.200

Mobley, E. M., Kunkel, B. N., and Keith, B. (1999). Identification, characterization and comparative analysis of a novel chorismate mutase gene in arabidopsis thaliana. *Gene* 240 (1), 115–123. doi: 10.1016/s0378-1119(99)00423-0

Park, S. Y., Vaghchhipawala, Z., Vasudevan, B., Lee, L. Y., Shen, Y. J., Singer, K., et al. (2015). Agrobacterium T-DNA integration into the plant genome can occur without the activity of key non-homologous end-joining proteins. *Plant J.* 81 (6), 934–946. doi: 10.1111/tpj.12779

Parkin, I. A., Koh, C., Tang, H., Robinson, S. J., Kagale, S., Clarke, W. E., et al. (2014). Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid brassica oleracea. *Genome Biol.* 15 (6), R77. doi: 10.1186/gb-2014-15-6-r77

Peňázová, E., Kopta, T., Jurica, M., Pečenka, J., Eichmeier, A., and Pokluda, R. (2018). Testing of inoculation methods and susceptibility testing of perspective cabbage breeding lines (Brassica oleracea convar. capitata) to the black rot disease caused by xanthomonas campestris pv. campestris. *Acta Universitatis Agriculturae Silviculturae Mendelianae Brunensis* 66 (1), 139–148. doi: 10.11118/actaun201866010139

Rifkin, S. A. (2012). *Quantitative trait loci (QTL): methods and protocols* (New York: Humana Press/Springer).

Saha, P., Kalia, P., Sharma, P., and Sharma, T. R. (2014a). Race-specific genetics of resistance to black rot disease [Xanthomonas campestris pv. campestris (Xcc) (Pammel) dowson] and the development of three random amplified polymorphic DNA markers in cauliflower. *J. Hortic. Sci. Biotechnol.* 89 (5), 480–486. doi: 10.1080/14620316.2014.11513109

Saha, P., Kalia, P., Sonah, H., and Sharma, T. R. (2014b). Molecular mapping of black rot resistance locus Xca1bo on chromosome 3 in Indian cauliflower (Brassica oleracea var. botrytis l.). *Plant Breed.* 133 (2), 268–274. doi: 10.1111/pbr.12152

Schaller, G. E., and Bleecker, A. B. (1995). Ethylene-binding sites generated in yeast expressing the arabidopsis ETR1 gene. *Science* 270 (5243), 1809–1811. doi: 10.1126/science.270.5243.1809

Sharma, B. B., Kalia, P., Yadava, D. K., Singh, D., and Sharma, T. R. (2016). Genetics and molecular mapping of black rot resistance locus Xca1bc on chromosome b-7 in Ethiopian mustard (Brassica carinata a. Braun). *PloS One* 11 (3), e0152290. doi: 10.1371/journal.pone.0152290

Staub, T. (1972). Factors influencing black rot lesion development in resistant and susceptible cabbage. *Phytopathology* 62, 722–728. doi: 10.1094/Phyto-62-722

Tang, R., Feng, T., Sha, Q., and Zhang, S. (2009). A variable-sized sliding-window approach for genetic association studies via principal component analysis. *Ann. Hum. Genet.* 73 (Pt 6), 631–637. doi: 10.1111/j.1469-1809.2009.00543.x

Taylor, J. D., Conway, J., Roberts, S. J., Astley, D., and Vicente, J. G. (2002). Sources and origin of resistance to xanthomonas campestris pv. campestris in brassica genomes. *Phytopathology* 92 (1), 105–111. doi: 10.1094/Phyto.2002.92.1.105

Tonu, N. N., Doullah, M.-u., Shimizu, M., Karim, M. M., Kawanabe, T., Fujimoto, R., et al. (2013). Comparison of positions of QTLs conferring resistance to <i>Xanthomonas campestris</i> pv. <i>campestris</i> in <i>Brassica oleracea</i&gt. *Am. J. Plant Sci.* 04 (08), 11–20. doi: 10.4236/ajps.2013.48A002

Vicente, J. G., Conway, J., Roberts, S. J., and Taylor, J. D. (2001). Identification and origin of xanthomonas campestris pv. campestris races and related pathovars. *Phytopathology* 91 (5), 492–499. doi: 10.1094/Phyto.2001.91.5.492

Vicente, J. G., and Holub, E. B. (2013). Xanthomonas campestris pv. campestris (cause of black rot of crucifers) in the genomic era is still a worldwide threat to brassica crops. *Mol. Plant Pathol.* 14 (1), 2–18. doi: 10.1111/j.1364-3703.2012.00833.x

Westfall, C. S., Xu, A., and Jez, J. M. (2014). Structural evolution of differential amino acid effector regulation in plant chorismate mutases. *J. Biol. Chem.* 289 (41), 28619–28628. doi: 10.1074/jbc.M114.591123

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* 136 (4), 1457–1468. doi: 10.1093/genetics/136.4.1457

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 27 July 2023
DOI 10.3389/fpls.2023.1226072

OPEN ACCESS

EDITED BY
Patricio Hinrichsen,
Agricultural Research Institute (Chile), Chile

REVIEWED BY
Dusan Stanisavljevic,
of Field And Vegetable Crops Novi Sad (IFVCNS), Serbia
Hongjun Yong,
Chinese Academy of Agricultural Sciences (CAAS), China

*CORRESPONDENCE
Thomas Lübberstedt
thomasl@iastate.edu

RECEIVED 20 May 2023
ACCEPTED 10 July 2023
PUBLISHED 27 July 2023

CITATION
Ledesma A, Ribeiro FAS, Uberti A, Edwards J, Hearne S, Frei U and Lübberstedt T (2023) Molecular characterization of doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic (BSSS) maize population.
Front. Plant Sci. 14:1226072.
doi: 10.3389/fpls.2023.1226072

Molecular characterization of doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic (BSSS) maize population

Alejandro Ledesma[1], Fernando Augusto Sales Ribeiro[1], Alison Uberti[1], Jode Edwards[2], Sarah Hearne[3], Ursula Frei[1] and Thomas Lübberstedt[1]*

[1]Department of Agronomy, Iowa State University, Ames, IA, United States, [2]USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA, United States, [3]International Maize and Wheat Improvement Center (CIMMYT), El Batan, Texcoco, Mexico

Molecular characterization of a given set of maize germplasm could be useful for understanding the use of the assembled germplasm for further improvement in a breeding program, such as analyzing genetic diversity, selecting a parental line, assigning heterotic groups, creating a core set of germplasm and/or performing association analysis for traits of interest. In this study, we used single nucleotide polymorphism (SNP) markers to assess the genetic variability in a set of doubled haploid (DH) lines derived from the unselected Iowa Stiff Stalk Synthetic (BSSS) maize population, denoted as C0 (BSSS(R)C0), the seventeenth cycle of reciprocal recurrent selection in BSSS (BSSS(R)C17), denoted as C17 and the cross between BSSS(R)C0 and BSSS(R)C17 denoted as C0/C17. With the aim to explore if we have potentially lost diversity from C0 to C17 derived DH lines and observe whether useful genetic variation in C0 was left behind during the selection process since C0 could be a reservoir of genetic diversity that could be untapped using DH technology. Additionally, we quantify the contribution of the BSSS progenitors in each set of DH lines. The molecular characterization analysis confirmed the apparent separation and the loss of genetic variability from C0 to C17 through the recurrent selection process. Which was observed by the degree of differentiation between the C0_DHL versus C17_DHL groups by Wright's F-statistics (FST). Similarly for the population structure based on principal component analysis (PCA) revealed a clear separation among groups of DH lines. Some of the progenitors had a higher genetic contribution in C0 compared with C0/C17 and C17 derived DH lines. Although genetic drift can explain most of the genetic structure genome-wide, phenotypic data provide evidence that selection has altered favorable allele frequencies in the BSSS maize population through the reciprocal recurrent selection program.

KEYWORDS

zea mays L., diversity, genetic resources, homozygous lines, genetic diversity

Introduction

The maize Iowa Stiff Stalk Synthetic (BSSS) population has undergone recurrent selection since 1939. This population was developed by intermating 16 inbred lines selected for superior stalk quality (Sprague and Jenkins, 1943). The C0 base population was subjected to multiple cycles of recurrent selection. Currently, C19 is available. The BSSS maize population has been under recurrent selection for increased grain yield, low grain moisture at harvest and increased resistance to root and stalk lodging. Phenotypic and genotypic changes have been observed in this population (Messmer et al., 1991; Labate et al., 1999; Hagdorn et al., 2003; Edwards, 2011; Gerke et al., 2015), suggesting loss of genetic variability from C0 to more advanced cycles of selection. As noted by Ledesma (2020) when they evaluated the level of phenotypic diversity and identified significant SNPs by GWAS in different cycles of recurrent selection of the BSSS population based on doubled haploid (DH) lines. Alleles present in a heterogeneous population of heterozygous individuals can be fixed in homozygous and homogenous DH lines and part of the genetic diversity in a population can be harnessed for breeding by production of DH lines (Böhm et al., 2017). However, the success of this approach relies on the choice of promising populations and extensive characterization of the produced DH lines (Böhm et al., 2017). The combination of DH technology with high-throughput genotyping drives progress in major maize breeding programs today (Andorf et al., 2019) and has been applied in this study to understand the evolution and genotypic composition of different cycles of BSSS maize population.

Molecular markers like single nucleotide polymorphisms (SNPs) have proven to be valuable for the characterization of maize germplasm and their application becoming more feasible over the past two decades due to the availability of new, high density and affordable genotyping technologies (Lu et al., 2009). Useful measures of the quality of genetic markers' polymorphisms are the expected heterozygosity (H_{exp}). Expected heterozygosity (H_e) is defined as the probability that any two alleles at a single locus, chosen randomly from the population, are different from each other (Nei and Roychoudhury, 1974; Nei, 1978).

The genetic relationship based on genetic distance was first defined by Nei (1973) as the difference between two samples that can be described by allelic variation, meaning that genotypes with many similar genes have a smaller genetic distance between them. The degree of genetic differentiation using the fixation index (F_{ST}) is a standard measure for the degree of genetic differentiation among subpopulations (Wright, 1951). The F_{ST} provides important insights into the evolutionary processes that influence the structure of genetic variation within and among populations (Holsinger and Weir, 2009). The F_{ST} estimates can identify regions of the genome that have been targeted for selection (Beckett et al., 2017; Guo et al., 2021; Wijayasekara and Ali, 2021). The comparison of F_{ST} from different genome regions can provide insights into populations demographic history (Holsinger and Weir, 2009).

Characterizing and understanding the genetic diversity and relationships of lines within a breeding program is essential for germplasm improvement (Andorf et al., 2019). Molecular markers have been used to estimate the relative strengths of evolutionary forces: mutation, natural selection, migration and genetic drift (Ouborg et al., 1999) and a possible loss of genetic diversity in specific populations, including BSSS (Gerke et al., 2015). Gerke et al. (2015), when evaluating different cycles of selection in a recurrent selection program, found that the populations steadily decreased in genetic diversity within populations and increased in genetic differentiation between populations mainly due to genetic drift and selection. According to the same authors, the C0 population has drifted away from the BSSS founders, despite the absence of intentional selection during the creation and maintenance of C0. In our study, we used different methods proposed as genetic diversity and differentiation measures using genotypic information. Additionally, we used developed DH lines instead of individual heterozygous plants representing the different cycles of the BSSS population.

Population structure is referred to as any form of relatedness among subgroups within the overall sample, including ancestry differences or cryptic relatedness (Sul et al., 2018). Population structure analysis involves grouping of individuals into subpopulations based on shared genetic variants and can be assessed through principal component analysis (PCA). PCA can identify differences in ancestry among populations and individuals, regardless of the historical patterns underlying population structure (Price et al., 2006; Zhu and Yu, 2009), since PCA clusters individuals based on the number of markers that are identical by state among them. Based on this grouping and relationship information among individuals, plant breeders can direct crosses, avoiding the mating closely related individuals and providing a reduction in inbreeding in their breeding programs. For instance, kinship coefficients have been used to estimate the genetic relationships within populations and to estimate the genetic contribution of a set of parents to its descendants (Yang et al., 2011; Ertiro et al., 2017; Wegary et al., 2019). Therefore, the estimation of kinship coefficients represents a way to utilize breeding resources more efficiently (Beckett et al., 2017).

An identity by descent (IBD) segment refers to DNA segments descended from common ancestors and could be useful to estimate the genetic relationships in a population. IBD occurs when identical alleles are inherited from a common ancestor and constitutes a measure of the degree of relationship between individuals (Wright, 1922). The estimation of the degree of the relationship depends on the description of an ancestral population, which by definition, is assumed to be the base from where past ancestry is no longer accounted (Wright, 1922). With the advent of high-throughput genotyping technologies, IBD segments can be estimated at a molecular scale. The identification of shared segments in the genome and haplotype information has been used for a range of purposes, including the quantification of inbreeding (Keller et al., 2011), identification of patterns of inheritance (Kirin et al., 2010), genotype imputation and haplotype inference (Browning and Browning, 2007), genetic characterization and diversity analysis (Nelson et al., 2008), the genetic contribution of a set of founder lines in commercial maize breeding programs (Coffman et al., 2019), and to improve the accuracy of genome-wide association analysis (GWAS; Maldonado et al., 2019) and genomic prediction (Won et al., 2020).

In this study of the BSSS, we propose to determine how much of the genomic variation in C0 has been lost during the selection process. C0 may be a reservoir of untapped favorable genetic diversity for previously unselected traits. Developing DH lines from earlier cycles of selection could be an alternative approach to conventional breeding for introduction of diversity into related elite lines. Genetic heterogeneity and high genetic load present in C0 could be overcome by production of DH lines (Böhm et al., 2017) to unlock genetic diversity. Diversity may have been lost not only due to selection can also be attributed to genetic drift or genetic hitchhiking effects, since no new genetic material was intentionally introduced into the BSSS population.

Our overall question in this and a companion paper Ledesma (2020) was whether potentially useful genetic diversity is available in earlier cycles of selection in the recurrent selection process, which may be more accessible sources of alleles compared with founding non-adapted landraces and other such genetic resources. Here, we used SNP markers to i) estimate and compare the genetic diversity within different subsets of DH lines derived from the BSSS maize population after different cycles of selection, ii) determine, if genetic diversity was lost from C0 to C17, iii) assess the genetic relationships and genetic divergence within and among the cycles of selection, and iv) perform a haplotype analysis based on IBD segments to quantify the contribution of the progenitors to each set of DH lines.

Materials and methods

Breeding populations

Three synthetic populations BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17 representing different cycles of selection in the reciprocal recurrent selection program with BSSS, and the Iowa Corn Borer Synthetic number 1 (BSCB1) were used to develop DH lines. The synthetic BSSS corresponds to the unselected base population (C0) formed by intermating 16 inbred lines selected for above average stalk quality in 1934 (Sprague, 1946). The C0 seed used came from subsequent cycles of seed multiplication in C0 for maintenance over time. The BSSS(R)C17 (C17) population corresponds to the seventeenth cycle of reciprocal recurrent selection with BSCB1 (Penny and Eberhart, 1971; Lamkey, 1992; Keeratinijakal and Lamkey, 1993; Edwards, 2011). Finally, BSSS/BSSS(R)17 was created by crossing plants from BSSS with plants in BSSS(R)C17 and intermating to create the BSSS/BSSS(R)C17 population (C0/C17). We also included in this study 14 (A3G-3-3-1-3, CI 540, I-159, IL12E, Oh 3167B, Os 420, Tr 9-1-1-6, WD 456, I224, LE 23, 461, Hy, AH83, CI 187-2) of the 16 known progenitors of the BSSS, plus the two parents (Fe and B2) of the F1B1 line. That is, a total of 16 progenitors were included in the study. Seed from the progenitor lines CI 617 and F1B1 were not available.

DH line development

Randomly selected individuals within each population were pollinated with a maternal haploid inducer BHI301 (Almeida et al., 2020) in an isolation field to generate the haploid seed. Seed produced from these plants was screened and kernels expressing the *R-nj* marker gene in the endosperm, but not in the embryo, were classified as haploid kernels. The haploid seed was germinated in plug trays in the Department of Agronomy greenhouse. Once seedlings developed 2-3 leaves, a colchicine treatment was applied following the protocol used by the DH Facility at ISU (Vanous et al., 2017). Two days after the colchicine treatment, haploid seedlings were transplanted in the field at the Agricultural Engineering and Agronomy Research Farm, Boone, IA. At flowering stage, putative DH_0 plants shedding pollen were self-pollinated to produce DH_1 seed. Seed multiplication was performed during subsequent generations and lines were screened for uniformity and discarded if they were segregating or variable. In total, 132 DH lines from BSSS(R)C0 (C0_DHL), 185 DH lines from BSSS(R)17 (C17_ DHL), and 170 DH lines from BSSS(R)C0/BSSS(R)17 (C0/C17_DHL) were obtained. The DH lines were developed by the DH Facility at ISU (http://www.plantbreeding.iastate.edu/DHF/DHF.htm).

Genotyping and quality control

Genomic DNA was extracted from DH line seedlings established in a greenhouse. Leaf tissue samples from three plants per DH line were collected at the 3-4 leaf developmental stage, and DNA extraction was done using the standard International Maize and Wheat Improvement Center (CIMMYT) laboratory protocol (Warburton, 2005). Genotyping was carried out using the Diversity Arrays Technology sequencing (DArT-seq) method (Kilian et al., 2012) provided by the Genetic Analysis Service for Agriculture (SAGA) laboratory at CIMMYT. DArT-seq is a high-throughput, robust, reproducible, and cost-effective genotyping technology based on genome complexity reduction using a combination of tailored restriction enzymes, followed by multiplexed sequencing of resulting libraries to simultaneously assay thousands of markers across the genome (Sansaloni et al., 2011). Across the samples assessed a total of 51,418 SNP markers were generated, of these 32,929 SNP markers were successfully aligned to the B73 RefGen_v4 (Jiao et al., 2017). Monomorphic and multi-allelic markers were removed. Un-imputed data without filtering for minor allele frequency (MAF) were used for further analyses.

The inbred line B73 was used as technical control and was repeated in seven separate plates to verify assay reproducibility. The resulting SNP core set was 24,885 SNP markers corresponding to 487 DH lines (132 C0_DHLs, 170 C0/C17_DHLs, 185 C17_DHLs) and 15 progenitors). The progenitor CI 187-2 was omitted because of heterozygosity greater than 8.8% (not expected in inbred lines) and was removed from further analyses. After this point, only 15 progenitors with low heterozygosity were used in the study.

Genotypic data analysis

Minor allele frequency analysis for each locus across the genotypes was calculated using the 24,885 SNP markers with the function 'Geno summary' analysis tool in the software TASSEL

v.5.2.64 (Bradbury et al., 2007). The expected heterozygosity (H_{exp}) was calculated to quantify the genetic variation in the maize lines sampled. The expected heterozygosity is defined as the probability that two alleles randomly chosen from the test sample are different (Nei, 1978). The expected heterozygosity was calculated using the R package "Poppr" (Kamvar et al., 2014), with the following formula: $H_{exp} = (\frac{n}{n-1})1 - \sum_{i=1}^{k} p_i^2$, where p is the allele frequency at a given locus, which goes from i to k, and n is the number of observed alleles for each locus (Nei, 1978).

The computation of dissimilarity coefficients or Euclidean genetic distance (Gower and Legendre, 1986) between DH lines and progenitor groups was performed with the 24,885 SNP markers using the R package "Poppr" (Kamvar et al., 2014). The genetic distances were calculated based on the average genetic distance of all lines within each other group. Cluster analyses were performed to subdivide the three sets of DH lines and the progenitor group into genetic subgroups using the Unweighted Pair Group Method with Arithmetic mean (UPGMA). Finally, dendrograms were constructed based on genetic distances using the visualization software Interactive Tree of Life (iTOL; Letunic and Bork, 2019).

To assess the degree of genetic differentiation between the groups of DH lines and the progenitors, we used the Wright's F-statistics (F_{ST}) on a per locus basis using the methodology described by Weir and Cockerham (1984), which accounts for unequal population sizes and sampling variances since heterozygous loci are weighted by the number of alleles observed in each population. The R package "hierfstat" (Goudet, 2005) was used to obtain estimates of F_{ST}. The F_{ST} values can range from zero to one, where high F_{ST} values showed a considerable difference in the allele frequency among two populations.

The pairwise relative kinship for all 487 DH lines and the 15 progenitors was estimated based on the 24,885 SNP markers using the software TASSEL v.5.2.64 (Bradbury et al., 2007) using the centered_IBS method (Endelman and Jannink, 2012). The relative kinship reflects the approximate degree of identity between two given individuals over the average probability of identity between two random individuals (Yu et al., 2006). The pairwise relative kinship was used to measure the genetic resemblance among individuals. A relative kinship close to zero indicates no relationship, and values close to one indicate a close relationship. Marker-based kinship coefficients show the relationship among lines based on genotypic information and rely on the marker allele frequencies in the reference population, which in practice is not known (Wang, 2014). However, 15 of the 16 progenitors of BSSS are known. These estimates commonly use the sample of genotyped individuals as the reference population, resulting in estimates that two homologous genes within or between individuals are shared by descent (Wang, 2014). Marker-based estimation of kinship coefficients can result in negative values. Wang (2014) states that the kinship coefficient's negative values could be interpreted as a lower probability that two homologous alleles are shared by descent compared with the probability that two alleles are taken at random from the reference population.

The 487 DH lines and the 15 progenitors were known to belong to the four subpopulations BSSS(R)C0, BSSS(R)C17, BSSS(R)C0/C17 and the progenitor groups, respectively. To examine the overall population structure across all lines, we performed a principal component analysis (PCA). PCA analysis allows the classification of individuals into genetically similar groups. PCA relies on reducing dimensionality by using principal components to maximize genetic variability (Price et al., 2006). Each principal component will account for a percentage of the total genetic variance by grouping the individuals into clusters with similar genetic information. After reducing dimensionality, a linear regression model was fitted to each of the axes of variation, and the residuals were extracted to compute associations (Price et al., 2006). PCA avoids any prior information about individual ancestries, the population of origin, and assumptions about the data, handling genome-wide data for thousands of individuals (Paschou et al., 2007). PCA was performed using the software GAPIT v.3 (Lipka et al., 2012). Bayesian Information Criterion (BIC; Schwarz, 1978) was used to identify the optimal number of principal components by selecting the lowest BIC model. The principal component results were used to display the first two principal components in R software (R Core Team, 2021).

The average linkage disequilibrium (LD) decay among SNP markers for each chromosome was determined in each group of DH lines using the squared Pearson correlation coefficient (r^2) among alleles at two loci, for all possible combinations of alleles, and then weighting them according to the allele frequency. P-values were determined by a two-sided Fishers Exact test (Bradbury et al., 2007). The option "Full Matrix LD" on TASSEL v.5.2.64 was used to calculate LD for every combination of sites in the alignment (Bradbury et al., 2007). The resulting data were imported into R (R Core Team, 2021) to create LD decay plots and fit a smooth line using Hill and Weir expectations of r^2 among adjacent sites (Hill and Weir, 1988).

To quantify the progenitors genetic contributions to the different sets of DH lines, we used high-resolution detection of identity by descend (IBD) segments. An IBD segment refers to DNA segments descended from common ancestors. IBD occurs when identical alleles are inherited from a common ancestor and could be used to estimate the genetic contribution. Estimation of IBD segments with genotypic data allows the quantification of the proportion of the covered genome descended from each progenitor. For the genetic contribution and the average LD decay among SNP marker analysis, a different filtering process of the genotypic data was conducted to have the most reliable SNP markers and ensure genotype concordance. From the 32,929 SNP markers successfully called within the B73 RefGen_v4 (Jiao et al., 2017). SNP markers with missing information rate above 10%, duplicated and monomorphic markers were removed in TASSEL v.5.2.64 (Bradbury et al., 2007). Genotypes were phased and imputed by using Beagle v.5.1 (Browning et al., 2018). Physical distance for each marker was converted to genetic distance using a dense 0.2 cM resolution map (Ogut et al., 2015), with on average 1385.6 kb per cM. After completing filtering and quality control, the genotypic data file contained 10,344 SNP markers for each of the 502 genotypes (487 DH lines and 15 progenitors) covering 2102.7 Mb (1,517.5 cM) of the genome and with one marker per 203.2 kb on average. The SNP markers not included in an IBD segment were referred to as non-IBD markers, while those within the IBD segment were labeled with the progenitor sharing the segment. The proportion of the genome descended from a progenitor was calculated by dividing the total number of SNP markers classified

as IBD by the total number of polymorphic SNP markers. Regions in the genome (IBD segments) that have been inherited from the progenitor were identified with the identity by descent linkage disequilibrium (IBDLD) program v.3.38 (Han and Abney, 2011; Han and Abney, 2013). The IBDLD program uses a probabilistic approach with a hidden Markov model to estimate IBD segments in pairs of individuals. The IBDLD program further expresses the emission probability conditioned on the true genotype of *n* previous loci to account for linkage disequilibrium (Han and Abney, 2011). IBD segments were constrained for each pair of individuals to have a minimum length of 350 kb, have more than 10 SNP markers and SNP markers with an IBD probability above 70%. These parameters force the segment to be a long IBD segment, avoiding segments formed by an occasional genotyping error or missing genotype occurring in otherwise-unbroken segments that could underestimate IBD segments for each pair of individuals (McQuillan et al., 2008).

Results

The initial number of SNP markers in the DArT-seq data set was 51,418. A total of 32,929 SNP markers were successfully called within the B73 RefGen_v4 (Jiao et al., 2017). After removing monomorphic and multi-allelic markers, the final SNP marker data set included 24,885 SNPs distributed across the ten chromosomes. The SNP density varied among chromosomes ranged from 3,976 to 1,688 markers on chromosome 1 and 10, respectively (Table 1). Heterozygosity varied from 1.2% on chromosomes 2 and 7 to 1.6% on chromosome 9, with a mean value of 1.3% across the ten chromosomes. We found heterozygous loci among the DHLs which ranged from 0.40 (C17_DHL045) to 2.24% (C0/C17_DHL146; Supplemental Table S1).

TABLE 1 Genotypic data summary for the 24,885 SNP markers and the entire panel of DH lines derived from different BSSS selection cycles.

Chromosome number	Number of SNP markers	Heterozygosity rate (%)
1	3,976	1.3
2	2,978	1.3
3	2,721	1.2
4	2,453	1.5
5	2,798	1.3
6	1,929	1.4
7	2,203	1.2
8	2,105	1.5
9	2,034	1.6
10	1,688	1.4
Genome-wide	24,885	1.3

Molecular characterization analysis

The 24,885 SNP markers were polymorphic with a MAF greater than zero (Figure 1). The average MAF was 0.19, 0.16, 0.13 and 0.07 in the progenitor, C0_DHL, C0/C17_DHL and C17_DHL groups, respectively (Table 2). The highest expected heterozygosity was in the progenitor's group (H_{exp} = 0.28), followed by the C0_DHL group with H_{exp} = 0.21 (Table 2). The lowest expected heterozygosity value was observed in the C17_DHL group as expected. In comparison, the group C0/C17_DHL had an expected heterozygosity value of H_{exp} = 0.19. The MAF and expected heterozygosity values all ranked populations in the same order. Higher values in progenitor and C0_DHL group were expected, which represents higher allelic variation in relation to the C0/C17_DHLs and C17_DHL groups. These values in the C0/C17_DHL group (F1 cross) were according with the expectation and were predictable values since we knew the parent populations (C0_DHL and C17_DHL) values.

Genetic differentiation analysis

The greatest genetic distance was observed between the progenitor group and the C17_DHL group (0.18) and the smallest genetic distance was observed between C17_DHL and C0/C17_DHL groups (0.11; Table 3). The UPGMA method separated the different groups of DH lines and the progenitor group (Figures 2–4). We observed that the grouping of lines and progenitors followed their origin. That is, lines and progenitors within groups were more related than among groups. In addition, we found high genetic diversity among the DH lines (C0_DHL, C0/C17_DHL and C17_DHL) and progenitors of each group.

The lowest F_{ST} among the DH lines was observed between the progenitors and the C0_DHL group (0.15). The highest value was observed between progenitors and C17_DHL (0.50; Table 3). Manhattan plots showed the genetic differentiation among the different comparisons performed between the progenitors and the different groups of DH lines across the ten chromosomes, with similar patterns across chromosomes (Figures 5 , 6). F_{ST} values of 1 and closer to 1 were observed between the progenitor group and the C17_DHL group across the genome as expected, demonstrating a considerable differentiation.

In relation to the pairwise relative kinship distribution for the entire set of 487 maize DH lines and 15 progenitors, 53.2% of the kinship coefficient was equal to 0 (Figure 7). Whereas, 46.0% of the entire panel ranged between 0 and 0.4, and only 0.8% were greater than 0.5. Thus, most lines were either not or only distantly related to each other.

Based on PCA, DH lines developed from BSSS can be divided into three subgroups (Figure 8). The first two principal components explained 12.5% of the total SNP variation in the entire panel. Based on discriminant analysis of principal components (DAPC), we observed a clear grouping of the DH lines into the C0_DHL, C17_DHL andC0/C17_DHL. The progenitor lines were grouped within the C0_DHL cluster, as expected, since the combination of

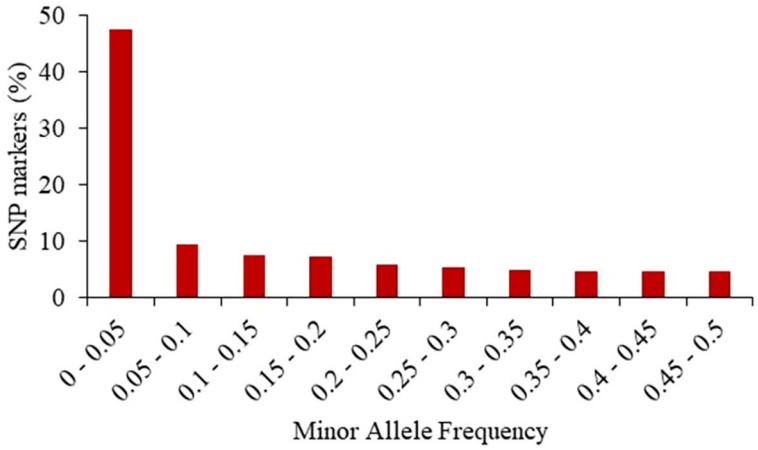


FIGURE 1
Frequency distribution of minor alleles in the entire panel of 487 BSSS DH lines and the 15 progenitors based on 24,885 SNP markers.

these 16 progenitor lines originated this population. The C0/C17_DHL group were scattered over a wider range, similar to the C0_DHL group.

The LD decay was variable across the ten chromosomes and different genetic regions within chromosomes in each group (Figure 9). The C17_DHL group showed the longest LD decay distances ranging from 1,229 to 2,709 kb on chromosomes 3 and 1, respectively. In contrast, the C0/C17_DHL group displayed the shortest LD decay distances (384 kb on chromosome 5 to 1,024 kb on chromosome 3). For C0_DHL, the LD decay varied from 486 kb to 1,322 kb for chromosomes 7 and 3, respectively.

For the progenitors' genetic contribution to each set of DH lines, a total of 10,344 polymorphic SNP markers distributed across the whole genome were used to estimate IBD segments among the 15 progenitors and 487 DH lines (Supplemental Table S2, Figure 10). In general, the progenitor A3G-3-3-1-3 had a low genetic contribution to the different sets of DH lines with 0.91, 0.87 and 0.63% in the C0_DHL, C0/C17_DHL and C17_DHL groups, respectively. In comparison, the progenitor WD 456 had a high genetic contribution to the different sets of DH lines with 5.76, 4.90 and 4.14% in the C0_DHL, C0/C17_DHL and C17_DH line groups, respectively. The progenitors CI 540 and Os 420 had a similar contribution to the different groups of DH lines. In general, the 15 progenitors evaluated had a higher genetic contribution to C0_DHLs, ranging from 0.91 to 5.87% for individual progenitors, compared with C0/C17_DHL (0.87 to 4.90%) and C17 (0.63 to 4.62%). The progenitor with the highest genetic contribution in C0 (Oh 3167B with 5.87%) had a lower contribution in C0/C17_DHL and C17 with 4.78 and 3.71%, respectively. On average, progenitor lines had 60.1% of the genome classified as identical by descent within C0_DHLs, 50.0% within the C0/C17_DHL and 41.6% within C17. The remaining 39.9, 50.0, and 58.4% in C0, C0/C17_DHL and

TABLE 2 Average Minor Allele Frequency (MAF) and expected heterozygosity (H_{exp}) within each group of DH lines and progenitors.

Group	Genotypes	Average MAF	H_{exp}
Progenitors	15	0.19 ± 0.001	0.28 ± 0.001
C0_DHL	132	0.16 ± 0.001	0.21 ± 0.001
C0/C17_DHL	170	0.13 ± 0.001	0.19 ± 0.001
C17_DHL	185	0.07 ± 0.001	0.09 ± 0.001

TABLE 3 Pairwise genetic distance and degree of genetic differentiation (F_{ST}) between different groups of DH lines and the progenitors of the BSSS maize population.

Group	Progenitors	C0_DHL	C0/C17_DHL	C17_DHL
Progenitors		0.168	0.170	0.175
C0_DHL	0.148		0.141	0.147
C0/C17_DHL	0.220	0.092		0.108
C17_DHL	0.496	0.340	0.131	

Below diagonal: pairwise F_{ST} estimates between different groups of DH lines and the progenitors, above diagonal: genetic distances.

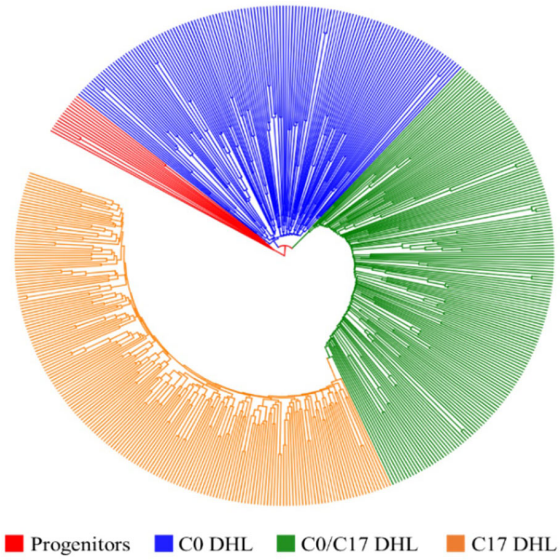


FIGURE 2
Dendrogram constructed from Euclidean genetic distance based on the UPGMA tree method for a panel of 15 progenitors and 495 DH lines derived from BSSS maize population.

C17, respectively are referred to as non-IBD markers. Those SNP markers were not included within the IBD segments between DH line groups and the progenitors.

Discussion

Molecular markers, including SNP markers have been used in many crops including maize for characterizing and quantifying genetic diversity of a given set germplasm for further improvement in a breeding program. The analysis of genetic variation among genetic materials is important to plant breeders, as it contributes to create a core set of germplasm, selecting parental lines, assigning heterotic groups, performs association analysis and prediction potential genetic gains for traits of interest. SNP markers, due to their abundance of availability of sophisticated, rapid, and affordable high-throughput detection systems, have become the principal resource for characterizing and quantifying genetic differences within and among species.

In the present study, the final SNP marker data set included 24,885 SNPs distributed across the ten chromosomes and 502 genotypes corresponding to DH lines derived from different cycles of recurrent selection (132 C0_DHL, 185 C17_ DHL, and 170 C0/C17_DHL) plus 15 progenitors of the BSSS maize population. The rationale of using un-imputed data without filtering for MAF was that the BSSS maize population came from 16 founder genotypes. For some SNP markers, an allele was provided by only one founder. The expected frequency would in such a case be ~6.2%. If genetic drift occurred, the actual frequency in C0 can be even lower. C0 seed used in this research came from subsequent cycles of seed multiplication for maintenance, increasing the chance of genetic drift to occur.

Changes in genetic diversity in different subsets of DH lines

When dividing the number of SNP with heterozygous loci by the total number of SNPs, we observed that our DH lines presented a very low rate of heterozygous loci (less than 3%). Therefore, our DH lines attained an appreciable level of homozygosity, and the DH technology was efficient to fix the loci without requiring further generations of purification. Higher MAF and expected heterozygosity values of the progenitor and C0_DHL groups (Table 2) were expected due to the large number of alleles that occurred in a few progenitor lines and were lost over recurrent selection cycles (Hagdorn et al., 2003). Additional recombination occurred because of population maintenance. Unfortunately, we do not have adequate records indicating how the seed has been maintained since 1939 when the population was created. Conversely, when comparing the C0_DHL and C17_DHL groups, we found a reduction in MAF and expected heterozygosity. The reduction in MAF among these groups was expected due to the recurrent selection process and genetic drift.

The high expected heterozygosity values found in the C0_DHL group were an indication for the presence of more rare alleles in C0. This could be an important source for new functional alleles of desirable traits, which have been lost during multiple generations of recurrent selection. Potential reduction in genetic diversity in advanced cycles were consistent with previous studies of the BSSS maize population in different cycles of the recurrent selection program (Messmer et al., 1991; Labate et al., 1997; Hagdorn et al., 2003; Hinze et al., 2005), where genome-wide genetic diversity has decreased across cycles of selection. Gerke et al. (2015) found a clear separation, when analyzing the progenitors and individuals from different cycles in the BSSS population. As this was a closed selection process, the substantial increase in genetic distance from

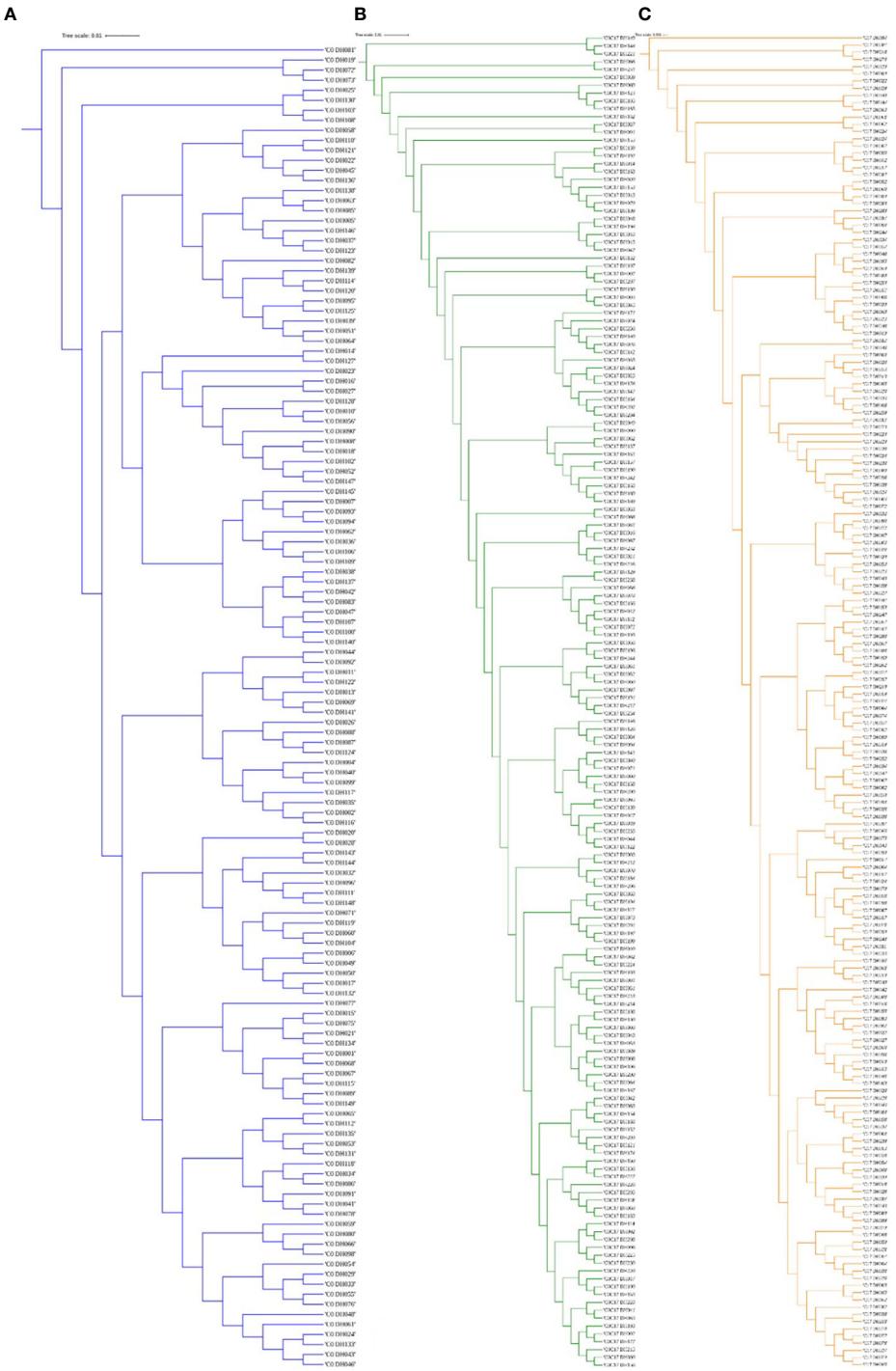


FIGURE 3
The dendrogram was constructed from Euclidean genetic distances based on the UPGMA tree method. **(A)** C0_DHL, **(B)** C0/C17_DHL and **(C)** C17_DHL of the BSSS maize population.

C0_DHL to C17_DHL could only arise from genetic differentiation due to selection and genetic drift (Gerke et al., 2015).

Improvement of plant characteristics like flag leaf angle, anthesis-silking interval, plant height, tassel branch number, total number of leaves and grain yield has been observed when advancing cycles in the BSSS recurrent selection program (Brekke et al., 2011; Edwards, 2011). These changes suggest fixation of favorable alleles during the recurrent selection program. Thus, exploring BSSS cycles using DH technology may reveal useful genetic diversity for plant characteristics left behind in the recurrent selection process and could be an important resource to help drive future genetic gains in maize breeding program.

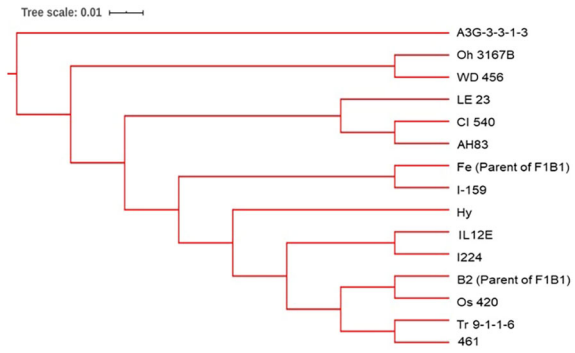


FIGURE 4
The dendrogram was constructed from Euclidean genetic distances based on the UPGMA tree method for the progenitors of the BSSS maize population.

Genetic relationship and divergence within and among cycles of selection

The Wright's F-statistics (F_{ST}) used to measure population substructure and the overall genetic divergence among the different groups showed that the degree of differentiation was higher between the progenitor inbred lines and the C17_DHL group compared to C0_DHL and C0/C17_DHL groups as expected since the two groups share fewer alleles. Lower F_{ST} values indicate limited differentiation between groups of DH lines. When we compare the F_{ST} values of C0_DHL versus C17_DHL, we observe a clear genetic differentiation among these two groups. These results can be confirmed with the wider genetic distance found among them, reflecting the uniqueness of most lines within these groups. Similar results were found by Gerke et al. (2015) when evaluating the progenitors and samples from different cycles of the BSSS maize population (C0, C4, C8, C12 and C16), indicating a clear differentiation between the founder lines and the population at C16 caused by the loss of different alleles within BSSS maize population. Gerke et al. (2015) conducted extensive simulations using BSSS founder haplotypes to gauge the roles of selection and drift among the cycles of selection and the results showed that most of the reduction in diversity observed among cycles can be attributed to genetic drift alone.

Population structure based on principal component analysis (PCA) is used to reveal genetic divergence among populations (Price et al., 2006). In this study, the results suggest a clear separation into three significant subgroups among all the BSSS DH lines and the progenitors. Also, we observed that the C0/C17_DHL group was scattered over a wide range, similar to C0_DHL, indicating a broader genetic divergence among these DH lines than for C17_DHL.

Kinship coefficients are defined by pedigree and can be estimated based on molecular information. Thus, it is possible to

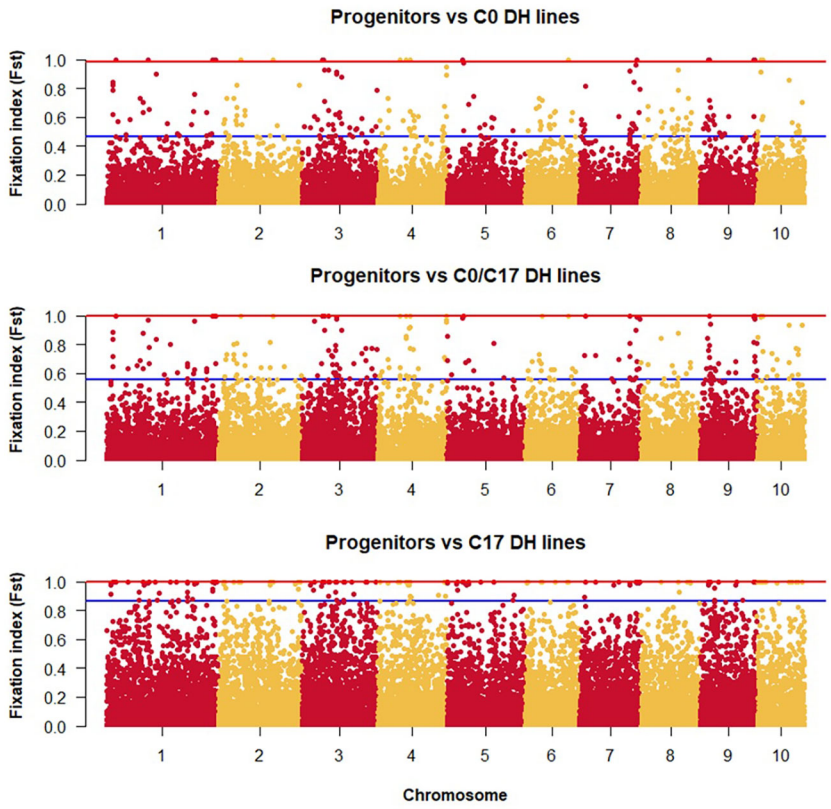


FIGURE 5
Genetic differentiation compares the progenitor group and the different groups of DH lines across chromosomes (x-axis) with the F_{ST} value (y-axis). Dots between the red and the blue lines represent the highest 1% of the F_{ST} values.

10.3389/fpls.2023.1226072

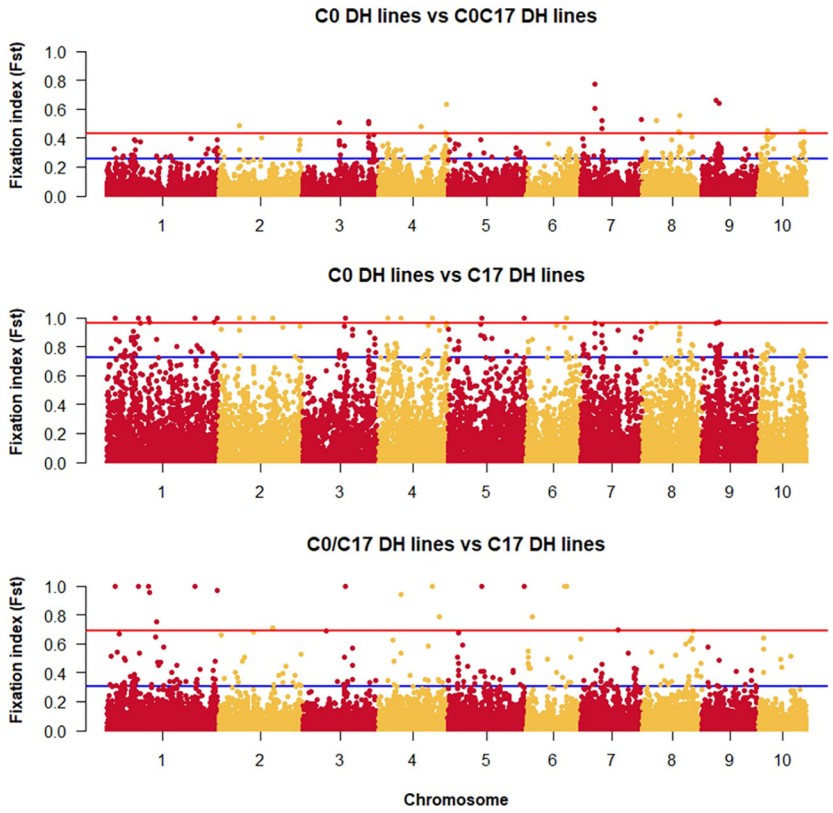


FIGURE 6
Genetic differentiation compares the different groups of DH lines across chromosomes (x-axis) with the F_{ST} value (y-axis). Dots between the rad and the blue lines represent the highest 1% of the F_{ST} values.

find hidden relationships. We found that most of the DH lines in the entire panel were distantly related to each other. Therefore, this shows us a low relationship between DH lines of the C17_DHL and C0_DHL. The estimation of the degree of the relationship depends on the description of an ancestral population, which by definition, is assumed to be the base from where the past ancestry is no longer accounted (Wright, 1922). Thus, the lower the number of generations separating the ancestral with the current population, the higher the kinship coefficient among individuals because of a reduced number of possible recombination events (Wang, 2014). Low or negative relative kinship coefficients among pairs of DH lines were found in the C0/C17_DHL group reflecting the uniqueness of most lines.

Linkage disequilibrium in BSSS DH lines

Linkage disequilibrium (LD) refers to the non-random co-segregation of alleles at two loci. Recombination events shuffle genetic material during meiosis among homologous chromosomes and cause LD to decay with increasing distance. Multiple factors are

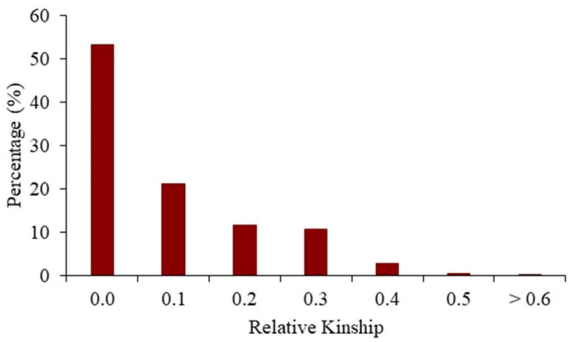


FIGURE 7
Distribution of pairwise relative kinship for 487 maize DH lines and 15 progenitors lines of the BSSS maize population calculated using 24,885 SNP markers.

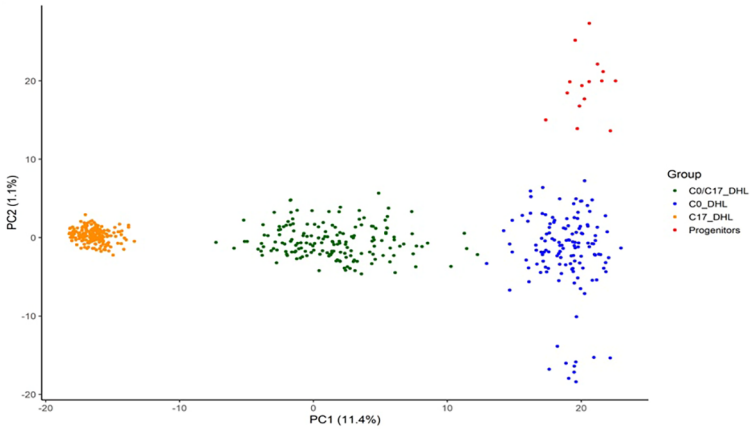


FIGURE 8
Scatter plot of the discriminant analysis of principal components based on 487 DH lines and 15 progenitors of the BSSS maize population. The dots represent each of the DH lines within their respective population. The axes represent the first two discriminant functions, respectively.

affecting LD in crops. Generally, LD decays faster in cross-pollinated crops, diverse populations, but also, different genes and genomic regions in the same crop can exhibit different rates of LD decay. It is expected in maize, for genome regions to decay at distances around 1 kb for exotic landraces, as described by (Romay et al., 2013). In the Ames panel subset corresponding to 384 lines (Pace et al., 2015) the LD decay rate was similar across chromosomes with an average distance of 10 kb throughout the genome. In this study, the LD decay distance among lines of the C17_DHL group was larger compared among lines of the C0_DHL and C0/C17_DHL groups. The longer LD decay distances in C17_DHL was consistent with the lower average MAF and expected heterozygosity results, as the rate of effective recombination declines over selection cycles due to the occurrence of bottlenecks or due to fixation for favorable alleles over time. The 17 cycles of recurrent selection did lead to a lower genetic diversity in the C17_DHL group, and LD decays more rapidly in pools of lines with higher genetic diversity (Romay et al., 2013; Wu et al., 2016). The distance over which LD persists determines the number and density of markers, and experimental design needed to perform an association analysis (Flint-Garcia et al., 2003). This was actually applied when generating the IBM Syn10 ultra-high-density map to precisely map a quantitative trait locus (Liu et al., 2015) at a higher genetic resolution than the IBM Syn4 map (Hu et al., 2016). In contrast, additional cycles of recurrent selection in the BSSS maize population increased homozygosity and LD decay distances due to selection and drift. Consistent with Gerke et al. (2015) genome-wide expected heterozygosity decreases steadily across cycles of selection. The loss of heterozygosity indicates the loss of different alleles within BSSS maize population.

Progenitor genetic contributions to different subsets of DH lines

On average, the mean genetic contribution of the BSSS progenitor lines estimated using high-resolution detection of IBD segments changed in the different groups of DH lines. The

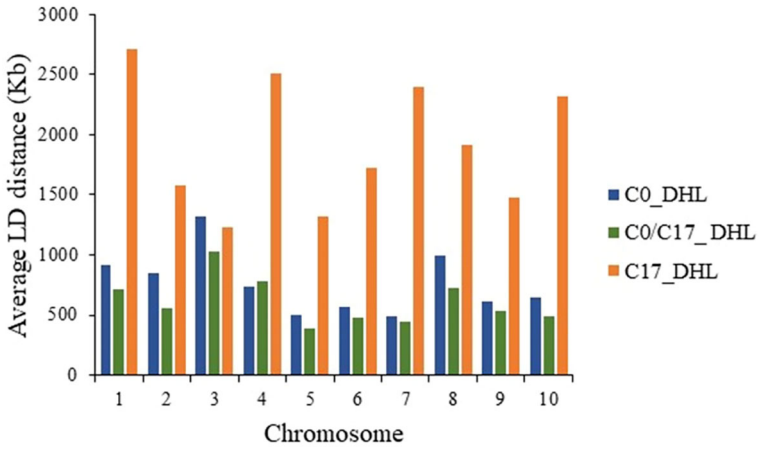


FIGURE 9
Linkage Disequilibrium (LD) decay distance per chromosome in the different groups of DH lines.

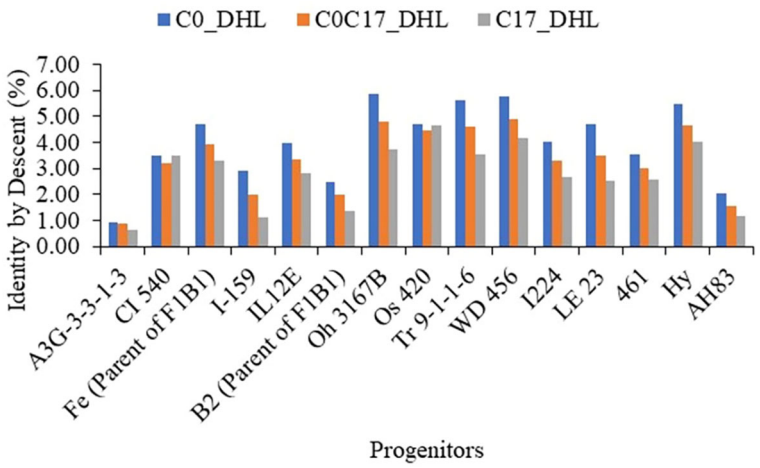


FIGURE 10
The genome's proportion classified as IBD among the BSSS progenitors inbred lines for each group of DH lines evaluated (C0_DHL, C0/C17_DHL and C17_DHL) identified with marker-based dissimilarity values.

progenitors had the highest genetic contribution in the C0_DHL group, due to their use in obtaining the population, and the lowest contribution in the C17_DHL group in relation to the other groups. This suggests that relationships caused by more recent ancestry had the most significant contribution in the IBD segments among individuals. Additionally, 17 cycles of recurrent selection have changed the allele frequencies in the C17_DHL group, because only individuals with superior performance for the selected traits contributed alleles to the next generation.

In the identification of regions in the genome inherited from the progenitors, we found prevalence of small to medium sized segments, where 50.4% of the segments were between 2.4 to 4.1 Mb. And, 28.2% of the segments ranged from 4.1 to 8.1 Mb inherited from the progenitor inbred lines. The number of segments decreased with increased length segments. IBD segment sizes from the progenitors changed across groups of DH lines. We found that some progenitors showed longer IBD segments in the C0_DHL group and others longer in the C17_DHL. Large, preserved regions in the genome could be associated with selection processes, resulting in long DNA segments inherited as a block from the parents. Therefore, under positive selection favoring a phenotype, a slight increase in LD surrounding the favored alleles will be produced. In these cases, the length of the IBD segment surrounding the alleles subject to selection will increase, experiencing less recombination at the population level (Albrechtsen et al., 2010). Albrechtsen et al. (2010) states that a reduced recombination rate in the genome, leading to significant LD, could be explained as a function of the effective population size. These could partially be explained by an increase in random genetic drift because of the population size, which will increase the length of DNA that will be shared among individuals in the population similar to what could happen in the C17_DHL with the 17 cycles of the recurrent selection process. The detection of long IBD segments in populations could be used as evidence for strong and recent selection processes because these segments have not suffered from recombination. However, many recombination's could have occurred because of subsequent cycles of seed multiplication and population maintenance. Unfortunately, we do not have adequate records indicating how the seed has been maintained since 1939 when the population was created. In cases where alleles within long IBD segments are in linkage disequilibrium, specifically in repulsion phase, unfavorable alleles will persist in the population, inducing the hitch-hiking effect and reducing the genetic diversity (Hospital and Chevalet, 1993). This hitch-hiking will increase genetic drift and significantly decrease the effective population size (Smith and Haigh, 1974). More studies should necessarily be done to confirm the possibility of the hitch-hiking effect having an effect in this population. Conversely, the restricted population size of both from founding (16 lines) and from continued population maintenance, may have provided the maintenance of long IBD segments. IBD segments shared between different groups of DH lines and the 15 progenitor lines will allow the estimation of genetic diversity and progenitor genetic contributions to new released lines.

In this study, we measured the genetic diversity among different sets of DH lines derived from the BSSS maize population and our results confirmed the separation from BSSS(R)C0 to BSSS(R)17 through the recurrent selection process. The selection process and the effective population size applied to the BSSS maize population have reduced the genetic variability. Consistent with previous studies (Messmer et al., 1991; Labate et al., 1997; Hagdorn et al., 2003; Hinze et al., 2005). Although genetic drift can explain most of the genetic structure genome-wide, phenotypic data provide evidence that selection has altered favorable allele frequencies in the BSSS maize population. We also found that the greatest genetic distance and F_{ST} observed between the progenitors group and the C17_DHL group demonstrated a clear genetic differentiation among groups caused by the loss of different alleles during the recurrent selection program in the BSSS maize population, reflecting the uniqueness of most lines within these groups of DH lines. Thus, these DH lines can be evaluated in replicated trials, and genomic selection can be applied for the estimation of the breeding value for each DH line. Additionally, DH lines derived from the BSSS maize population could be ideal for association mapping due to

the low population structure. Thus, we could identify genes or regions in the genome associated with a particular trait. Using genome-based data and DH technology was a powerful tool for access to the genetic diversity available in C0_DHL or C0/C17_DHL groups, which would be beneficial to incorporate in BSSS(R)17 to broaden its genetic variation while minimizing yield or other penalties. Thus, the results of this research will also help maize breeders to explore useful genetic variation for further improvement in a breeding program.

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: Iowa State University DataShare, accession 22893878. DOI: https://doi.org/10.25380/iastate.22893878.v1.

Author contributions

TL, AL, JE, UF and SH conceived and designed the experiments. AL analyzed the genotypic data and conducted the molecular characterization. AL and FA conducted the IBD analysis. AL, AU and TL wrote the manuscript, with contributions from all the other authors. All authors contributed to the article and submitted and approved the submitted section.

Funding

Funding for this work was provided by USDA's National Institute of Food and Agriculture (NIFA) Project, No. IOW04314, IOW01018, and IOW05510; and NIFA award 2018-51181-28419. Funding for this work was also provided by the R.F. Baker Center sfor Plant Breeding, Plant Sciences Institute, and K.J. Frey Chair in Agronomy at Iowa State University

Acknowledgments

Alejandro Ledesma Miramontes acknowledges the National Council for Science and Technology (CONACYT), International Maize and Wheat Improvement Center (CIMMYT) and the National Institute for Agricultural, Livestock, and Forestry Research (INIFAP) for the scholarship 2016 for Ph.D. studies.

Conflict of interest

The authors declare that they have no conflicts of interest.
Availability of data and material data transparency

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.2023.1226072/full#supplementary-material

References

Albrechtsen, A., Moltke, I., and Nielsen, R. (2010). Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186, 295–308. doi: 10.1534/genetics.110.113977

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

Andorf, C., Beavis, W. D., Hufford, M., Smith, S., Suza, W. P., Wang, K., et al. (2019). Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* 132, 817–849. doi: 10.1007/s00122-019-03306-3

Beckett, T. J., Morales, A. J., Koehler, K. L., and Rocheford, T. R. (2017). Genetic relatedness of previously Plant-Variety-Protected commercial maize inbreds. *PloS One* 12, 1–23. doi: 10.1371/journal.pone.0189277

Böhm, J., Schipprack, W., Utz, H. F., and Melchinger, A. E. (2017). Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. *Theor. Appl. Genet.* 130, 861–873. doi: 10.1007/s00122-017-2856-x

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. *J. Bioinform.* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Brekke, B., Edwards, J., and Knapp, A. (2011). Selection and adaptation to high plant density in the Iowa Stiff Stalk Synthetic maize (*Zea mays* L.) population. *Crop Sci.* 51, 1965–1972. doi: 10.2135/cropsci2010.09.0563

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Coffman, S. M., Hufford, M. B., Andorf, C. M., and Lübberstedt, T. (2019). Haplotype structure in commercial maize breeding programs in relation to key founder lines. *Theor. Appl. Genet.* 133, 547–561. doi: 10.1007/s00122-019-03486-y

Edwards, J. (2011). Changes in plant morphology in response to recurrent selection in the Iowa Stiff Stalk Synthetic maize population. *Crop Sci.* 51, 2352–2361. doi: 10.2135/cropsci2010.09.0564

Endelman, J. B., and Jannink, J. L. (2012). Shrinkage estimation of the realized relationship matrix. *Genes|Genomes|Genetics* 2, 1405–1413. doi: 10.1534/g3.112.004259

Ertiro, B. T., Semagn, K., Das, B., Olsen, M., Labuschagne, M., Worku, M., et al. (2017). Genetic variation and population structure of maize inbred lines adapted to the mid-altitude sub-humid maize agro-ecology of Ethiopia using single nucleotide polymorphic (SNP) markers. *BMC Genom.* 18, 1–11. doi: 10.1186/s12864-017-4173-9

Flint-Garcia, S. A., Thornsberry, J. M., and Buckler, E. S. B.IV (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Gerke, J. P., Edwards, J. W., Guill, K. E., Ross-Ibarra, J., and McMullen, M. D. (2015). The genomic impacts of drift and selection for hybrid performance in maize. *Genetics* 201, 1201–1211. doi: 10.1534/genetics.115.182410

Goudet, J. (2005). HIERFSTAT, a Package for R to compute and test hierarchical F-statistics. *Mol. Ecol. Notes* 2, 184–186. doi: 10.1111/j.1471-8278

Gower, J. C., and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3, 5–48. doi: 10.1007/BF01896809

Guo, R., Chen, J., Petroli, C. D., Pacheco, A., Zhang, X., San Vicente, F., et al. (2021). The genetic structure of CIMMYT and US inbreds and its implications for tropical maize breeding. *Crop Sci.* 61, 1666–1681. doi: 10.1002/csc2.20394

Hagdorn, S., Lamkey, K. R., Frisch, M., Guimaraes, P. E., and Melchinger, A. E. (2003). Molecular genetic diversity among progenitors and derived elite lines of BSSS and BSCB1 maize populations. *Crop Sci.* 43, 474–482. doi: 10.2135/cropsci2003.4740

Han, L., and Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype data. *Genet. Epidemiol.* 2335, 557–567. doi: 10.1002/gepi.20606

Han, L., and Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *Eur. J. Hum. Genet.* 21, 205–211. doi: 10.1038/ejhg.2012.148

Hill, W. G., and Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* 33, 54–78. doi: 10.1016/0040-5809(88)90004-4

Hinze, L. L., Kresovich, S., Nason, J. D., and Lamkey, K. R. (2005). Population genetic diversity in a maize reciprocal recurrent selection program. *Crop Sci.* 45, 2435–2442. doi: 10.2135/cropsci2004.0662

Holsinger, K. E., and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* 100, 639–650. doi: 10.1038/nrg2611

Hospital, F., and Chevalet, C. (1993). Effects of population size and linkage on optimal selection intensity. *Theor. Appl. Genet.* 86, 775–780. doi: 10.1007/BF00222669

Hu, S., Lübberstedt, T., Zhao, G., and Lee, M. (2016). QTL mapping of low-temperature germination ability in the maize IBM Syn4 RIL population. *PloS One* 11, 1–11. doi: 10.1371/journal.pone.0152795

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., et al. (2017). Improved maize reference genome with single-molecule technologies. *Nature* 546, 524–527. doi: 10.1038/nature22971

Kamvar, Z. N., Tabima, J. F., and Grüunwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2013, 1–14. doi: 10.7717/peerj.281

Keeratinijakal, V., and Lamkey, K. R. (1993). Responses to reciprocal recurrent selection in BSSS and BSCB1 maize populations. *Crop Sci.* 33, 73–77. doi: 10.2135/cropsci1993.0011183X003300010012xh

Keller, M. C., Visscher, P. M., and Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics* 189, 237–249. doi: 10.1534/genetics.111.130922

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). "Diversity arrays technology: a generic genome profiling technology on open platforms,"In: Pompanon, F., Bonin, A. (eds) Data Production and Analysis in Population Genomics. *Methods Mol. Biol* (Hertfordshire, UK: Humana Press). 67–89. doi: 10.1007/978-1-61779-870-2_5

Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., Mckeigue, P. M., and Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PloS One* 5, 1–7. doi: 10.1371/journal.pone.0013996

Labate, J. A., Lamkey, R., Lee, M., and Woodman, W. L. (1997). Molecular genetic diversity after reciprocal recurrent selection in BSSS and BSCBI maize populations. *Crop Sci.* 37, 416–423. doi: 10.2135/cropsci1997.0011183X003700020018x

Labate, J. A., Lamkey, K. R., Lee, M., and Woodman, W. L. (1999). Temporal changes in allele frequencies in two reciprocally selected maize populations. *Theor. Appl. Genet.* 99, 1166–1178. doi: 10.1007/s001220051321

Lamkey, K. (1992). Fifty years of recurrent selection in the Iowa stiff stalk synthetic maize population. *Maydica* 37, 19–28.

Ledesma, A. (2020). Molecular and phenotypic characterization of doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic maize population [Dissertation/PhD thesis]. [Ames, (IA)]: Iowa State University.

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, 256–259. doi: 10.1093/nar/gkz239

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

Liu, H., Niu, Y., Gonzalez-Portilla, P. J., Zhou, H., Wang, L., Zuo, T., et al. (2015). An ultra-high-density map as a community resource for discerning the genetic basis of quantitative traits in maize. *BMC Genom.* 16, 1–16. doi: 10.1186/s12864-015-2242-5

Lu, Y., Yan, J., Guimarães, C. T., Taba, S., Hao, Z., Gao, S., et al. (2009). Molecular characterization of global maize breeding germplasm based on genome-wide single nucleotide polymorphisms. *Theor. Appl. Genet.* 120, 93–115. doi: 10.1007/s00122-009-1162-7

Maldonado, C., Mora, F., Scapim, C. A., and Coan, M. (2019). Genome-wide haplotype-based association analysis of key traits of plant lodging and architecture of maize identifies major determinants for leaf angle: hapLA4. *PloS One* 14, e0212925. doi: 10.1371/journal.pone.0212925

McQuillan, R., Leutenegger, A. L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., et al. (2008). Runs of homozygosity in European populations. *Am. J. Hum. Genet.* 83, 359–372. doi: 10.1016/j.ajhg.2008.08.007

Messmer, M. M., Melchinger, A. E., Lee, M., Woodman, W. L., Lee, E. A., and Lamkey, K. R. (1991). Genetic diversity among progenitors and elite lines from the Iowa Stiff Stalk Synthetic (BSSS) maize population: comparison of allozyme and RFLP data. *Theor. Appl. Genet.* 83, 97–107. doi: 10.1007/BF00229231

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.* 70, 3321–3323. doi: 10.1073/pnas.70.12.3321

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583–590. doi: 10.1093/genetics/89.3.583

Nei, M., and Roychoudhury, A. (1974). Sampling variances of heterozygosity and genetic distance. *Genetics* 76, 379–390. doi: 10.1093/genetics/76.2.379

Nelson, P. T., Coles, N. D., Holland, J. B., Bubeck, D. M., Smith, S., and Goodman, M. M. (2008). Molecular characterization of maize inbreds with expired U.S. plant variety protection. *Crop Sci.* 48, 1673–1685. doi: 10.2135/cropsci2008.02.0092

Ogut, F., Bian, Y., Bradbury, P. J., and Holland, J. B. (2015). Joint-multiple family linkage analysis predicts within-family variation better than single-family analysis of the maize nested association mapping population. *Heredity* 114, 552–563. doi: 10.1038/hdy.2014.123

Ouborg, N. J., Piquot, Y., and Van Groenendael, J. M. (1999). Population genetics, molecular markers and the study of dispersal in plants. *J. Ecol.* 87, 551–568. doi: 10.1046/j.1365-2745.1999.00389.x

Pace, J., Gardner, C., Romay, C., Ganapathysubramanian, B., and Lübberstedt, T. (2015). Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). *BMC Genom.* 16, 1–12. doi: 10.1186/s12864-015-1226-9

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., et al. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PloS Genet.* 3, 1672–1686. doi: 10.1371/journal.pgen.0030160

Penny, L. T., and Eberhart, S. A. (1971). Twenty years of reciprocal recurrent selection with two synthetic varieties of maize (*Zea mays* L.). *Crop Sci.* 11, 900–903. doi: 10.2135/cropsci1971.0011183X001100060041x

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, 904–909. doi: 10.1038/ng1847

R Core Team (2021). "R: A language and environment for statistical computing," (R Foundation for Statistical Computing).

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14, 1–18. doi: 10.1186/gb-2013-14-6-r55

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., et al. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. *BMC Proc.* 5, 1–2. doi: 10.1186/1753-6561-5-s7-p54

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Smith, M. J., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35. doi: 10.1017/S0016672308009579

Sprague, G. F. (1946). Early testing of inbred lines of corn. *J. Am. Soc Agron.* 38, 108–117. doi: 10.2134/agronj1946.00021962003800020002x

Sprague, G. F., and Jenkins, M. T. (1943). A comparison of synthetic varieties, multiple crosses, and double crosses in corn. *J. Agron.* 35, 137–147. doi: 10.2134/agronj1943.00021962003500020007x

Sul, J. H., Martin, L. S., and Eskin, E. (2018). Population structure in genetic studies: Confounding factors and mixed models. *PloS Genet.* 14, 1–22. doi: 10.1371/journal.pgen.1007309

Vanous, K., Vanous, A., Frei, U. K., and Lübberstedt, T. (2017). Generation of maize (*Zea mays*) doubled haploids via traditional methods. *Curr. Protoc.* 2, 147–157. doi: 10.1002/cppb.20050

Wang, J. (2014). Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. *J. Evol. Biol.* 27, 518–530. doi: 10.1111/jeb.12315

Warburton, M. (2005). Laboratory Protocols: CIMMYT applied molecular genetics laboratory, 3rd ed. (Mexico, D.F: CIMMYT).

Wegary, D., Teklewold, A., Prasanna, B. M., Ertiro, B. T., Alachiotis, N., Negera, D., et al. (2019). Molecular diversity and selective sweeps in maize inbred lines adapted to African highlands. *Sci. Rep.* 9, 1–15. doi: 10.1038/s41598-019-49861-z

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370. doi: 10.2307/2408641

Wijayasekara, D., and Ali, A. (2021). Evolutionary study of maize dwarf mosaic virus using nearly complete genome sequences acquired by next-generation sequencing. *Sci. Rep.* 11, 1–14. doi: 10.1038/s41598-021-98299-9

Won, S., Park, J. E., Son, J. H., Lee, S. H., Park, B. H., Park, M., et al. (2020). Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00134

Wright, S. (1922). Coefficients of inbreeding and relationship. *Am. Nat.* 56, 330–338.

Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.* 15, 323–354.

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., et al. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* 129, 753–765. doi: 10.1007/s00122-016-2664-8

Yang, X., Gao, S., Xu, S., Zhang, Z., Prasanna, B. M., Li, L., et al. (2011). Characterization of a global germplasm collection and its potential utilization for analysis of complex quantitative traits in maize. *Mol. Breed.* 28, 511–526. doi: 10.1007/s11032-010-9500-7

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., 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, 203–208. doi: 10.1038/ng1702

Zhu, C., and Yu, J. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* 182, 875–888. doi: 10.1534/genetics.108.098863

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 October 2023
DOI 10.3389/fpls.2023.1270166

OPEN ACCESS

EDITED BY
Patricio Hinrichsen,
Agricultural Research Institute, Chile

REVIEWED BY
Manje S. Gowda,
The International Maize and Wheat Improvement Center (CIMMYT), Kenya
Sivakumar Sukumaran,
The University of Queensland, Australia

*CORRESPONDENCE
Thomas Lübberstedt
thomasl@iastate.edu

†PRESENT ADDRESS
Darlene L. Sanchez,
Texas A&M AgriLife Research, Beaumont, TX, United States

RECEIVED 31 July 2023
ACCEPTED 18 September 2023
PUBLISHED 09 October 2023

CITATION
Sanchez DL, Santana AS, Morais PIC, Peterlini E, De La Fuente G, Castellano MJ, Blanco M and Lübberstedt T (2023) Phenotypic and genome-wide association analyses for nitrogen use efficiency related traits in maize (*Zea mays* L.) exotic introgression lines.
Front. Plant Sci. 14:1270166.
doi: 10.3389/fpls.2023.1270166

Phenotypic and genome-wide association analyses for nitrogen use efficiency related traits in maize (*Zea mays* L.) exotic introgression lines

Darlene L. Sanchez[1†], Alice Silva Santana[1], Palloma Indiara Caproni Morais[1], Edicarlos Peterlini[2], Gerald De La Fuente[1], Michael J. Castellano[1], Michael Blanco[1,3] and Thomas Lübberstedt[1*]

[1]Department of Agronomy, Iowa State University, Ames, IA, United States, [2]Department of Agronomy, State University of Maringá, Maringá, PR, Brazil, [3]Department of Agriculture, Agricultural Research Service (USDA-ARS), Ames, IA, United States

Nitrogen (N) limits crop production, yet more than half of N fertilizer inputs are lost to the environment. Developing maize hybrids with improved N use efficiency can help minimize N losses and in turn reduce adverse ecological, economical, and health consequences. This study aimed to identify single nucleotide polymorphisms (SNPs) associated with agronomic traits (plant height, grain yield, and anthesis to silking interval) under high and low N conditions. A genome-wide association study (GWAS) was conducted using 181 doubled haploid (DH) lines derived from crosses between landraces from the Germplasm Enhancement of Maize (BGEM lines) project and two inbreds, PHB47 and PHZ51. These DH lines were genotyped using 62,077 SNP markers. The same lines from the *per se* trials were used as parental lines for the testcross field trials. Plant height, anthesis to silking interval, and grain yield were collected from high and low N conditions in three environments for both *per se* and testcross trials. We used three GWAS models, namely, general linear model (GLM), mixed linear model (MLM), and Fixed and Random model Circulating Probability Unification (FarmCPU) model. We observed significant genetic variation among the DH lines and their derived testcrosses. Interestingly, some testcrosses of exotic introgression lines were superior under high and low N conditions compared to the check hybrid, PHB47/PHZ51. We detected multiple SNPs associated with agronomic traits under high and low N, some of which co-localized with gene models associated with stress response and N metabolism. The BGEM panel is, thus, a promising source of allelic diversity for genes controlling agronomic traits under different N conditions.

KEYWORDS

candidate gene, quantitative trait locus, diversity, genetic resources, abiotic stress

1 Introduction

Nitrogen (N) is critical to promote crop growth and development and to increase grain yield. In cereals such as maize, the application of N fertilizers is an essential agronomic practice (Nag and Das, 2022). Although N fertilizer markedly improves the yield of maize, its excessive use often leads to run-off, which causes the eutrophication of rivers and other bodies of water (Wani et al., 2021). In this context, more than half of the N fertilizer applied to maize is lost to the environment (Ladha et al., 2016; Yu et al., 2022). As an example, N leaching from maize-based cropping systems is the primary cause of hypoxia in the Gulf of Mexico (Goolsby et al., 2000; Alexander et al., 2007). Hence, it is increasingly important to screen genotypes for N use efficiency (NUE) and explore those that have higher NUE and are better suited to N limitation.

Improving NUE in maize would not only help to reduce N fertilization in the field but may also increase productivity in N-deficient environments. However, NUE is a complex trait in which interactions between genetic and environmental factors are involved. Traits such as anthesis-silking interval, plant height, and grain yield have the potential to be used as parameters for NUE screening, since they play an essential role in N acquisition and N utilization in maize, the two main components of NUE (Gheith et al., 2022). NUE-related traits have been successfully used in maize (Kumari et al., 2021), rice Liu et al. (2016b), and potatoes (Getahun et al., 2020) to identify genotypes with better performance under low N conditions. In addition, studies combining quantitative genetics and molecular markers support a strategy of great potential for plant breeders to analyze the genetic architecture of complex traits such those related to NUE. In this context, genome-wide association studies (GWAS) have been widely used to capture complex trait variation down to the genome level by exploring both historical and evolutionary recombination events in maize (Verzegnazzi et al., 2021; Ma et al., 2022; Wu et al., 2022; Xu et al., 2023).

In US elite germplasm, only a small fraction of the total available genetic diversity in maize (<10 out of 300 maize races) is currently used (Andorf et al., 2019). The Germplasm Enhancement in Maize (GEM) project of United States Department of Agriculture—Agricultural Research Service (USDA-ARS) has the objective of improving maize productivity by broadening the genetic base of commercial maize cultivars through evaluating, identifying, and introducing useful genes from maize landraces (Pollak, 2003; Salhuana and Pollak, 2006). In the allelic diversity component of the GEM project, doubled haploid (DH) lines were derived from BC1F1 or F1 crosses between tropical and subtropical accessions and elite inbreds PHB47 (stiff stalk) and PHZ51 (non-stiff stalk), which are expired plant variety protection (ex-PVP) lines (Brenner et al., 2012), to enable photoperiod adaptation of these materials to Midwest US conditions. Currently, the released DH lines are known as BGEM lines, where B indicates Iowa State University, the place where the DH lines were developed (Vanous et al., 2018).

In this study, BGEM lines *per se*, and their testcrosses, were evaluated in field trials under low and high (normal) N conditions for agronomic traits related to NUE. GWAS analyses for the agronomic traits under low (LN) and high N (HN) conditions were conducted. The main objective was to identify novel alleles associated with agronomic traits under low N conditions, which can aid in improving NUE in maize. The specific objectives were to (i) determine the extent of variation of agronomic traits for the BGEM panel grown under HN and LN conditions, (ii) establish correlations among the agronomic traits, (iii) identify associations between SNP markers and agronomic traits grown under HN and LN conditions, and (iv) evaluate the co-localization of these SNPs with putative candidate genes and/or previously identified QTL for traits related to NUE in the inbred and testcross populations.

2 Materials and methods

2.1 Plant materials

In total, 66 GEM accessions from Central and South America were crossed with the expired PVP lines PHB47 and PHZ51. Most of the F_1 seeds were backcrossed once with PHB47 and PHZ51, respectively, to produce the BC_1F_1 generation as described in Sanchez et al. (2018). A total of 181 BGEM lines and inbred lines PHB47 and PHZ51 were used in *per se* field trials. The DH lines were produced using the protocol described by Vanous et al. (2017), wherein BC_1F_1 or F_1-derived crosses between GEM accessions and PHB47 or PHZ51 were crossed with the inducer hybrid RWS 9 × RWK-76 (Röber et al., 2005) to produce haploid seed, which was identified based on the *R-nj* color marker Liu et al. (2016a). In the subsequent planting season, putative haploids were grown in the greenhouse, where colchicine treatment was applied to seedlings at the three to four leaf developmental stage to promote genome doubling. Haploid plants were transplanted in the field and self-pollinated to produce DH lines. Seed of these lines was increased at the USDA North-Central Region Plant Introduction Station in Ames, Iowa during the summer of 2013 and at the Iowa State University Agricultural Engineering and Agronomy Farm in 2014. In total, 74 and 105 DH lines were obtained from the crosses with the recurrent parents PHZ51 (non-stiff stalk) and PHB47 (stiff stalk), respectively.

The same lines from the *per se* trials were used as parental lines for the testcross field trials. They were divided according to heterotic group membership (i.e., stiff-stalk and non-stiff stalk), and each group was planted in separate isolation plots in Ames during the summer of 2014. Two rows and two ranges of pollen parent surrounded each isolation plot. Inside, for every two rows of female, there was one row of male. There were three replications or rows of each DH line, randomly distributed per isolation plot. In one isolation plot, all lines belonging to the stiff-stalk group (e.g., DH lines with PHB47 as recurrent parent) that were used as female parents were detasseled before anthesis, and PHZ51 was used as pollen parent. In the other isolation plot, all non-stiff stalk lines (e.g., DH lines with PHZ51 as recurrent parent) were detasseled and crossed with PHB47. In total, 74 and 105 testcrosses obtained from the cross with PHZ51 and PHB47, respectively, were evaluated.

10.3389/fpls.2023.1270166

2.2 Field trials

In this study, a combination of location and year was considered as an environment. Within each environment, two N conditions were evaluated: HN and LN. No fertilizer was applied within the LN condition in all environments. For the HN condition, 261.60 kg N ha^{-1} was applied in the form of 32% urea–ammonium nitrate (UAN) fertilizer before planting via liquid broadcast and immediately incorporated with tillage. Three environments were used for the *per se* trials: at Iowa State University Agricultural Engineering and Agronomy Farm (42.0204° latitude, −93.7738° longitude, 335 m elevation) in Ames, IA, during the summers of 2014 (Ames 2014) and 2015 (Ames 2015), and at the Iowa State University Northeast Research and Demonstration Farm (42.93811° latitude, −92.57018° longitude, 317.742 m elevation) in Nashua, IA, during the summer of 2015 (Nashua 2015).

Two environments were used for the testcross trials, which were performed at the same farms from Ames and Nashua during the summer of 2015. No N fertilizer was applied to the Nashua LN location in 2014, and oats were planted in that area before, in order to deplete the soil N content. For the testcross evaluation in Ames 2015, two LN locations were used. One has historically been planted with maize, and no fertilizer has been applied in that location for several years. The other LN location in Ames 2015 did not receive any fertilizer treatment and was planted with non-nodulating soybeans in the previous year (2014). Therefore, for the testcrosses trials, the maize–maize location was referred to as Ames 2015A, and the soybean–maize location was referred to as Ames 2015B. Only one HN location was used for testcrosses trials in Ames 2015 environment.

Soil samples were collected right before sowing, and the samples were analyzed in the Ames trial plots in 2015. Using a probe, 10 samples per location were collected in the top 30 cm of the soil at randomly selected areas, and samples for each trial were bulked, thoroughly mixed, and submitted to the ISU Soil and Plant Analysis Laboratory at the Department of Agronomy to determine total N and carbon (C) content (McGeehan and Naylor, 1988). The results of samples were collected and analyzed in the Ames trial plots in 2015 (Table 1). Results reported C and N as the percentage (%) of C or N in the dried sample (g C or N per 100 g sample). For logistical issues, it was not possible to collect soil samples in Nashua.

All trials were planted following a randomized complete block design (RCBD), in two-row plots. Two ranges of filler were planted at the front and back and four rows at the left and right sides of each trial. Each row was 5.64 m long, and the rows were spaced 0.76 m apart. Planting density was 65,323 plants ha^{-1}.

2.3 Agronomic traits evaluated

Plant height (PHT) and grain yield (GY) were measured in all trials, while anthesis to silking interval (ASI) data were only collected at the Ames trials. ASI was calculated using the difference in growing degree units (GDUs) between anthesis and silking times. Days to anthesis was recorded as the number of days from sowing to the day when 50% of the plants in the plot had anthers extruded outside the glumes. Days to silking were recorded as the number of days from sowing to the day when 50% of the plants in the plot had silks emerging from the ears. Days to anthesis and silking were converted to growing degree units (GDUs), which were calculated according to the following equation: $GDUs = \frac{T_{max}+T_{min}}{2}$, where T_{max} is the maximum daily temperature which is set to 30°C when T_{max} exceed 30°C, and T_{min} is the minimum temperature and is set to 10°C when T_{min} falls below 10°C. PHT in centimeters was taken from the ground surface to the topmost end of the central tassel spike. GY was obtained from two-row plots using a harvesting combine, where grain weight and moisture content were measured. Yield in tons per hectare was computed after moisture content was adjusted to 15.50%.

2.4 Statistical analysis of agronomic traits

Data analysis was performed separately for the *per se* and testcross trials fitting the following linear model: $Y_{ijkl} = \mu + E_i + R(E)_{ij} + N_l + EN_{il} + G_k + EG_{ik} + NG_{lk} + ENG_{ikl} + \varepsilon_{ijkl}$, where Y_{ijkl} is the observation in the k^{th} genotype in the j^{th} replication in the i^{th} environment and l^{th} N rate; μ is the overall mean; E_i is the effect of the i^{th} environment; $R(E)_{ij}$ is the effect of j^{th} replication nested within the i^{th} environment; N_l is the effect of the l^{th} N rate; EN_{il} is the interaction effect of the i^{th} environment and l^{th} N rate; G_k is the effect of the k^{th} genotype; EG_{ik} is the effect of the interaction of the i^{th} environment with the k^{th} genotype; EG_{lk} is the effect of the interaction of the l^{th} N rate with the k^{th} genotype; ENG_{ikl} is the effect of the interaction of the i^{th} environment and l^{th} N rate with the k^{th} genotype; and ε_{ijkl} is the residual error.

The procedure PROC MIXED from the software package SAS (SAS Institute Inc., North Carolina, USA) was used to perform the analysis of mixed model, where N rate was fixed, and the other factors were random. Variance components, $\sigma_g^2, \sigma_{g\times e}^2, \sigma_e^2$, were estimated accordingly, where $\sigma_g^2, \sigma_{g\times e}^2, \sigma_e^2$ correspond to the genotypic variance, genotype by environment interaction variance, and error variance, respectively. Broad-sense heritability

TABLE 1 Results of soil samples collected and analyzed in the Ames trial plots.

Condition	Trial	N (%)	C (%)	Location
High N	*Per se*	0.39	6.30	Ames
	Testcross	0.35	4.10	Ames
Low N	*Per se*	0.16	2.02	Ames
	Testcross	0.17	2.09	Ames 2015A
	Testcross	0.17	2.00	Ames 2015B

(h^2) on an entry mean basis for each trait under each N condition and in the combined analysis were estimated as follows (Hallauer et al., 2010): $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_e^2}{r}}$ and $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{g\times e}^2}{n} + \frac{\sigma_e^2}{rn}}$, where r is the number of replications within each environment, and n is the number of environments.

For each N condition, best linear unbiased predictions (BLUPs) from all inbred lines and testcrosses across the environments were estimated for all measurements. This was also implemented using PROC MIXED in SAS 9.3 (SAS Institute Inc., 2011). The BLUPs from the combined analysis within each N condition were used to calculate Pearson correlations among traits using PROC CORR function in SAS 9.3 (SAS Institute Inc., 2011).

2.5 Molecular marker data

The BGEM lines were genotyped using 955,690 genotyping-by-sequencing (GBS) markers (Elshire et al., 2011). GBS data were generated at the Cornell Institute for Genomic Diversity (IGD) laboratory. After filtering out markers with more than 25% missing data, below 2.5% minor allele frequency, and monomorphic markers, 247,775 markers were left for further analyses. For markers at the same genetic position (0 cM distance), only one marker was randomly selected. The final number of markers used for further analyses was 62,077 markers distributed across all 10 chromosomes.

The average number of recombination events per line was substantially greater than expected. Therefore, the genotypic data were corrected for monomorphic markers that were located between flanking markers displaying donor parent genotypes. The correction was based on Bayes theorem, with an underlying assumption that very short distances of a marker with recurrent parent (RP) genotype to flanking markers with donor genotype are more likely due to identity of marker alleles for that particular SNP between RP and donor, instead of a rare double recombination event. These short RP segments interspersed within donor segments were tested for the null hypothesis that a double recombination occurred and were either corrected or kept as original genotype, accordingly, based on p-values from the Bayes theorem (Vanous et al., 2018). After correction, the donor genome composition was closer to the expected 25%, compared to the original marker data, and the average number of recombination events was substantially reduced (Sanchez et al., 2018). Genotype data of the testcrosses were generated using the "create hybrid genotypes" function in TASSEL 5.2.61 (Bradbury et al., 2007) with genotype information from the BGEM lines *per se*, PHB47 and PHZ51.

2.6 Genome-wide association studies

BLUPs from the combined analysis of the traits ASI, PHT, and GY for HN and LN conditions, in the *per se* and testcross trials, were used for GWAS. In order to balance false-positives and false-negatives in detecting significantly associated SNPs, three statistical models were implemented, namely, (1) General Linear Model (GLM) + PCA (Q), where the PCA output from GAPIT was used as a covariate to account for fixed effects due to population structure; (2) Mixed Linear Model (MLM; Yu et al., 2006), where PCA and kinship (K) were used as covariates; and (3) FarmCPU (Fixed and random model Circulating Probability Unification), where Q was also used as covariate, but has additional algorithms to solve the confounding problems between testing markers and covariates Liu X. et al. (2016). The R package GAPIT (Lipka et al., 2012) was used to conduct GWAS for all three models. Additive genetic model was implemented when performing GWAS for *per se* trials, while dominant genetic model was used for the testcross trials.

Multiple testing in GWAS was accounted for using the statistical program simpleM (Gao et al., 2010; Johnson et al., 2010), which calculates the number of informative SNPs (Meff_G) using R statistical software (R Core Team, 2014). First, a correlation matrix for all markers was constructed, and the corresponding eigenvalues for each SNP locus were calculated. GAPIT (Lipka et al., 2012) was then used to calculate a composite linkage disequilibrium (CLD) correlation directly from the SNP genotypes, and once this SNP matrix was obtained, *Meff_G* was calculated, and this value was used to compute for the multiple testing threshold in the same way as the Bonferroni correction method, where the significance threshold (α=0.05) was divided by the M_{eff_G} (α/M_{eff_G}). For this study, based on the α level of 0.05, the multiple testing threshold level was set at 8.10×10^{-7}.

The available maize genome sequence (B73; RefGen_v4) was used as the reference genome for candidate gene identification. Candidate genes were identified using the Ensembl Biomart tool (Kinsella et al., 2011). Genes were considered as candidates if a significantly associated SNP marker with phenotypic variance explained (PVE) higher than 10% was located within the range of linkage disequilibrium (LD) decay observed for each chromosome (upstream and downstream). Candidate genes corresponding to each SNP were checked according to the SNP marker's physical position in the MaizeGDB molecular marker database (http://www.maizegdb.org; Portwood et al., 2019). Functional annotations of candidate genes were predicted in NCBI (http://www.ncbi.nlm.nih.gov/gene) and were also compared to previously published candidate genes.

3 Results

3.1 Field performance of BGEM lines *per se* under high and low nitrogen conditions

According to the soil chemical analysis (Table 1), the N content at LN trials was considerably lower than at HN trials, indicating that the N-depleting effort had been successful in reducing N levels. In addition, all measured traits were affected by N conditions, and most of them had their means reduced by the N deficiency (Table 2). We observed wide ranges on the tested traits under LN and HN (Table 2). However, the N stress negatively affected the genotypic variation among the DH lines, and the ranges were much larger under HN than under LN for almost all traits, except for ASI in Ames 2014. For this trait, the range under LN was equal to

10.3389/fpls.2023.1270166

TABLE 2 Summary statistics of agronomic traits in BGEM lines *per se* and testcrosses grown under different N conditions.

Environment	Trait	Low N				High N				Mixed models analysis		
		Mean	Max	Min	H^2	Mean	Max	Min	H^2	$\hat{\sigma}^2_G$	$\hat{\Phi}_N$	$\hat{\sigma}^2_{NG}$
BGEM lines *per se*												
Ames 2014	ASI	25.00	85.45	−19.53	0.57	10.74	67.59	−25.08	0.51	1,237.76**	95,296.22**	232.06 *
	PHT	197.27	240.33	159.90	0.75	222.04	272.04	158.60	0.82	379.13**	128,974.98**	5.25*
	GY	2.18	3.92	0.91	0.29	2.92	5.85	0.66	0.81	0.76**	75.41**	0.26**
Ames 2015	ASI	37.85	83.06	11.73	0.28	19.62	106.78	14.84	0.61	1,365.54**	209,157.88**	40.97ns
	PHT	180.59	215.00	141.66	0.58	226.85	289.74	181.96	0.63	368.49**	397,964.28**	22.29ns
	GY	1.03	2.05	0.71	0.24	2.37	4.68	1.18	0.42	0.29**	246.94**	0.25**
Nashua 2015	PHT	228.38	276.13	168.60	0.78	240.41	296.12	182.96	0.83	22.14**	27,976.98**	7.84ns
	GY	3.01	5.13	1.21	0.66	4.21	6.96	1.38	0.69	1.30**	182.83**	0.32**
Combined	ASI	31.44	99.34	−4.68	0.42	15.21	95.69	−16.97	0.30	35.50**	270,250.54**	10.27**
	PHT	202.10	245.65	159.56	0.61	229.78	281.77	177.27	0.59	402.81**	519,339.69**	0.43ns
	GY	2.10	3.44	1.19	0.21	3.17	5.55	0.97	0.40	0.61**	319.50**	0.22**
Testcrosses												
Ames 2015A	ASI	19.91	43.29	10.05	0.24	−0.14	23.29	−11.25	0.04	238.23**	25,7926.92**	23.30ns
	PHT	248.75	278.38	207.56	0.41	334.44	379.48	284.45	0.68	232.32**	8,763.92**	41.46*
	GY	3.48	4.68	2.35	0.30	8.27	12.00	3.98	0.62	0.44**	1,890.28**	1.05**
Ames 2015B	ASI	19.84	35.89	12.67	0.25	–	–	–	–	149.44**	228,403.25**	53.88**
	PHT	259.71	285.67	220.41	0.60	–	–	–	–	246.01**	52,119.82**	24.95*
	GY	4.35	5.95	2.32	0.56	–	–	–	–	0.97**	1,559.56**	0.68**
Nashua 2015	PHT	297.44	322.20	248.66	0.72	318.69	346.89	266.96	0.58	202.87**	105,567.88**	138.23ns
	GY	8.29	8.86	7.72	0.09	11.11	16.11	6.39	0.37	0.73**	3,535.22**	0.46**
Combined	ASI	19.87	46.20	8.58	0.23	−0.14	23.17	−11.19	0.53	125.32**	231,387.94**	114.81**
	PHT	268.64	296.37	217.35	0.49	326.57	367.56	269.18	0.69	219.66ns	39,175.14**	4.23ns
	GY	5.37	6.37	3.95	0.28	8.28	11.76	4.93	0.52	0.74**	3,427.98**	0.83ns

[a]ASI, anthesis to silking interval (GDU); PHT, plant height (cm); GY, Grain Yield (t ha^{-1}), H^2, broad-sense heritability; $\hat{\sigma}^2_G$, genotypic variance component estimate; $\hat{\Phi}_N$, quadratic component of nitrogen fixed effect; $\hat{\sigma}^2_{NG}$, variance component of nitrogen rate by genotype interaction; *significant at p = 0.05; **significant at p = 0.01; ns, not significant.
(-) means data were not collected.

104.98, while under HN, it was equal to 92.67. On the other hand, traits such as PHT presented wider ranges under HN conditions. In Ames 2014, PHT ranged from 158.60 cm to 272.04 cm under HN and from 159.90 cm to 240.33 cm under LN. In Ames 2015, the same trait had a ranged from 181.96 cm to 289.74 cm under HN and from 141.66 cm to 215.00 cm under LN. In general, higher values of standard deviation (SD) were also observed under HN conditions. For example, ASI had SD equal to 18.36 under HN in Ames 2015, while under LN, it was equal to 12.34. In Ames 2014, PHT had SD equal to 20.36 under HN and 16.25 under LN.

PHT and GY were affected by N deficiency and had their means reduced under LN conditions (Table 2). While the mean of PHT under HN was equal to 222.04, it was equal to 197.27 cm under LN condition in Ames 2014. In Ames 2015, GY had a mean of 2.37 t ha^{-1} under HN, while under LN, it was equal to 1.03 t ha^{-1}. On the other hand, ASI had higher means under LN than under HN condition. In Ames 2014 and Ames 2015, ASI had means equal to 10.74 and 19.62 under HN, respectively, and equal to 25.00 and 37.85 under LN, respectively. We observed that GY was the trait most negatively affected by N deficiency and presented the highest mean reduction in response to the LN across all environments. The decrease in the mean under LN compared to HN was equal to 25.34%, 56.54%, 28.50%, and 33.75% in Ames 2014, Ames 2015, and Nashua 2015 and in the combined analysis, respectively.

Variance components due to genotype were highly significant ($p< 0.01$) by the likelihood ratio test for all traits in the *per se* trials (Table 2). In addition, variance components due to genotypes × N rates interaction were highly significant ($p< 0.05$) for almost all traits. In general, the heritability estimates were higher under HN than under LN conditions. For example, GY heritability estimate under HN in

Ames 2014 was equal to 0.81, while under LN condition, it was equal to 0.29. In Ames 2015, ASI had heritability equal to 0.61 under HN and equal to 0.28 under LN condition. In the combined analysis, the heritability estimates were low to intermediate (<0.70). In this context, GY had the lowest heritability estimate under LN condition (0.21) and the intermediate one under HN (0.40). Across all environments, the highest yielding BGEM lines under LN were BGEM-0137-S, BGEM-0044-S, BGEM-0127-N, and BGEM-0243-S with GY ranging from 3.12 t ha^{-1} to 3.44 t ha^{-1}, and 52 out of the 179 BGEM lines performed better than PHB47 (GY = 2.41 t ha^{-1}). On the other extreme, DH lines BGEM-0223-N, BGEM-0225-N, BGEM-0247-N, BGEM-0237-N, and BGEM-0165-S performed poorly with yields ranging from 1.19 t ha^{-1} to 1.30 t ha^{-1}.

3.2 Performance of testcrosses under high and low nitrogen conditions

Similar to the *per se* trials, the ranges were much larger under HN than under LN for almost all traits, except for ASI in the combined analysis (Table 2). This difference was even more pronounced with GY. In Ames 2015A, GY values ranged from 3.98 t ha^{-1} to 12.00 t ha^{-1} under HN, while under LN, it ranged from 2.35 t ha^{-1} to 4.68 t ha^{-1}. In Nashua 2015, GY ranged from 6.39 t ha^{-1} to 16.11 t ha^{-1} under HN and from 7.72 t ha^{-1} to 8.86 t ha^{-1} under LN condition. In general, SD values were also higher under HN conditions, except for ASI in Ames 2015A and in the combined analysis. For GY in Ames 2015A, the SD was equal to 1.43 and 0.39 under HN and LN conditions, respectively. PHT and GY were affected by N conditions, and their means reduced with the N deficiency. The percentage of reduction in the mean was stronger for GY. The GY reduction mean was equal to 57.92%, 25.38%, and 35.14% in Ames 2015A, Nashua 2015, and in the combined analysis, respectively (Table 2). Conversely, ASI increased its means under LN condition. In Ames 2015A, ASI means were equal to −0.14 and 19.91 under HN and LN conditions, respectively.

The statistical analysis conducted within environment for testcrosses showed that, for almost all traits, there was significant effect of genotype ($p< 0.01$), except for PHT in the combined analysis. Variance components due to genotypes × N rates interaction were highly significant ($p< 0.01$) for GY in all environments, while for ASI and PHT, the significance depended on the environment where they were evaluated. In relation to the heritability estimates within environments, we observed that PHT had the highest estimates among the three traits, ranging from 0.41 to 0.72 under LN and from 0.58 to 0.68 under HN. The heritability estimates for GY ranged from 0.09 to 0.56 under LN and from 0.37 to 0.62 under HN (Table 2). In general, heritability estimates in the testcross trials across environments were higher under HN than under LN. For example, in the combined analysis of ASI, heritability estimates under HN were equal to 0.53 and 0.23 under LN.

Testcrosses performing best under LN across environments were BGEM-0258-S/PHZ51, BGEM-0112-S/PHZ51, BGEM-0070-S/PHZ51, BGEM-0115-S/PHZ51, BGEM-0233-S/PHZ51, and BGEM-235-N/PHB47, with yields ranging from 6.13 t ha^{-1} to 6.33 t ha^{-1}. The lowest yields were obtained for BGEM-0166-S/PHZ51, BGEM-0263-S/PHZ51, BGEM-0269-S/PHZ51, BGEM-0078-S/PHZ51, and BGEM-00129-N/PHB47, ranging from 3.95 t ha^{-1} to 4.19 t ha^{-1}. GY of the checks, PHB47/PHZ51 and its reciprocal PHZ51/PHB47, under LN were 6.37 t ha^{-1} and 5.85 t ha^{-1}, respectively. Testcrosses outperforming the GY of PHB47/PHZ51 were identified in the Ames environments. In Ames 2015B environment, there were testcrosses that outperformed PHB47/PHZ51, with BGEM-0112-S/PHZ51, BGEM-0155-S/PHZ51, and BGEM-0226-S/PHZ51 performing better than PHB47/PHZ51 under both LN and HN. The testcrosses BGEM-0001-N/PHB47, BGEM-0044-S/PHZ51, BGEM-0111-S/PHZ51, BGEM-0114-S/PHZ51, and BGEM-0115-S/PHZ51 performed consistently better than PHB47/PHZ51 under the two LN environments in Ames.

3.3 Correlations among and within *per se* and testcross agronomic traits

Within BGEM lines *per se*, significant and close positive correlations were observed for PHT evaluated under different N conditions (r = 0.91), and GY (r = 0.69) and ASI (r = 0.75; Table 3). Moderate negative correlations were observed between ASI and GY under HN (r = −0.50) and LN (r = −0.48). Within the testcross, a high positive correlation was observed between PHT under HN and PHT under LN condition (r = 0.78) and between GY and PHT under LN (r = 0.66). ASI under HN was not significantly correlated with neither GY under HN and LN nor with PHT under LN. There were also no significant correlations observed between ASI under LN and PHT under HN and GY under HN (Table 3). In addition, there was no strong correlation (r > 0.60) between GY with the other two traits neither under HN nor under LN for BGEM lines and their testcrosses. Therefore, according to our results, we could not use PHT and ASI as indirect selectors for GY.

Weak to moderate (r< 0.60) correlation coefficients were observed between the performance of testcross and *per se* genotypes (Table 4). The highest correlation coefficients were observed between testcross PHT under HN and *per se* lines PHT under HN (r = 0.52) and LN (r = 0.52). Testcross PHT under LN also correlated well with *per se* PHT under both HN (r = 0.49) and LN (r = 0.52). According to the correlation coefficients, there is no possibility to use any trait from the *per se* performance to predict the performance of testcross hybrids under neither N condition.

3.4 Genome-wide association studies for agronomic traits in *per se* and testcross trials

To reduce the impact of environmental variability, BLUP values across the three environments (Ames 2015A, Ames 2015B and Nashua 2015) were used for association study. No SNPs were found when performing GWAS with MLM model. A total of seven significant SNPs were found by applying FarmCPU and GLM models (Table 5; Figure 1). The same SNPs detected by FarmCPU were detected by GLM. This result indicates that these

TABLE 3 Pearson correlation of agronomic traits in BGEM lines and testcrosses grown under low nitrogen (LN) and high nitrogen (HN) conditions across environments.

Trait		HN			LN		
		GY[a]	PHT	ASI	GY	PHT	ASI
HN	GY		0.36**	−0.50**	0.69**	0.29**	−0.42**
	PHT	0.43**		−0.15*	0.14*	0.91**	−0.04^{ns}
	ASI	−0.12^{ns}	−0.23**		−0.46**	−0.08^{ns}	0.75**
LN	GY	0.48**	0.49**	−0.11^{ns}		0.18**	−0.48**
	PHT	0.40**	0.78**	−0.11^{ns}	0.66**		0.03^{ns}
	ASI	−0.07^{ns}	−0.14^{ns}	0.43**	−0.27**	−0.21**	

Values above the diagonal are correlations among BGEM lines per se, and values below the diagonal are correlations among testcrosses.
[a]GY, Grain Yield (t ha^{-1}); PHT, plant height (cm); ASI, anthesis to silking interval (GDU); *significant at p = 0.05; **significant at p = 0.01; ^{ns}not significant.

common SNPs have high reliability. For simplicity, we presented the results from FarmCPU, and the subsequent analysis mainly focused on those seven SNPs.

For the *per se* data, the GWAS analysis identified significant SNPs only for ASI under HN condition. Interestingly, one of the two SNPs (S2_190189512) had PVE >30%. This SNP is within the gene model GRMZM2G414252, located between 190,556,326 and 190,557,054 bp on Chromosome 2. The SNP marker S1_13685600 ($P = 1.11\times10^{-11}$, SNP effect = 7.58) was also significantly associated with ASI under HN conditions. The associated gene model GRMZM2G037912 (14,081,196–14,083,562 bp in Chromosome 1) was identified as a putative vesicle-associated membrane protein (Table 6).

For testcross data, three significant SNP markers each were found for PHT under both LN and HN, but did not overlap. None of the SNPs affected more than one trait. The SNP marker S1_104874404 on Chromosome 1 was significantly associated with PHT under LN ($P = 2.49\times10^{-9}$, SNP effect = 8.80) with a PVE equal to 24.8%. This SNP is located within the gene model GRMZM2G158976 (105,553,409–105,554,335 bp), and encodes a VQ motif-containing protein. GRMZM2G070271 is 308,612 bp away from GRMZM2G158976 and encodes a xyloglucan endotransglucosylase/hydrolase protein. For PHT under HN conditions, one SNP marker had a PVE higher than 10% (S3_179633217). This SNP marker is within the gene model GRMZM2G087619 (177,609,579–177,634,652 bp), identified as sister chromatid cohesion protein on Chromosome 3. S2_209927372 was significantly associated with GY under HN ($p = 5.44\times10^{-7}$, SNP effect = −0.71). It is worth noting that this SNP marker explained more than 40% of phenotypic variance. The gene model GRMZM2G311187 (209,688,288–209,689,726) co-locates with this SNP, which encodes for a phosphatase protein. Other putative gene models identified by significant associations are listed in Table 6.

4 Discussion

4.1 Effect of nitrogen deficiency on agronomic traits

Screening maize genotypes for yield-related traits tested under LN conditions and optimal-N conditions is critical for long-term maize production in areas with low N fertility. In our study, we evaluated a panel of BGEM lines and their respective testcrosses. Information about population structure, genetic diversity, and linkage disequilibrium of BGEM lines have been reported (Sanchez et al., 2018; Ma et al., 2020; Zuffo et al., 2022). We

TABLE 4 Correlations of agronomic traits between BGEM lines *per se* and testcrosses grown under different Nitrogen (N) conditions across environments.

Per se traits		Testcross traits					
		High N			Low N		
		GY[a]	PHT	ASI	GY	PHT	ASI
High N	GY	0.11^{ns}	0.14*	−0.16*	0.17*	0.08^{ns}	−0.13^{ns}
	PHT	0.14*	0.52**	−0.07^{ns}	0.15*	0.49**	0.01^{ns}
	ASI	0.00^{ns}	−0.05^{ns}	0.28**	−0.03^{ns}	0.01^{ns}	0.21**
Low N	GY	0.03^{ns}	0.05^{ns}	−0.17*	0.19**	0.03^{ns}	−0.22**
	PHT	0.12^{ns}	0.52**	−0.05^{ns}	0.18**	0.52**	0.00^{ns}
	ASI	0.03^{ns}	0.06^{ns}	0.26**	0.01^{ns}	0.13^{ns}	0.23**

[a]GY, Grain Yield (t ha^{-1}); PHT, Plant height (cm); ASI, Anthesis to silking interval (GDU); *significant at p=0.05; **significant at p=0.01; ^{ns}not significant.

10.3389/fpls.2023.1270166

TABLE 5 Significant SNP markers information associated with agronomic traits of BGEM lines *per se*, and their testcrosses, grown under high nitrogen (HN) and low nitrogen (LN) conditions.

Trait	SNP	Chr	P-value	Effect	MAF	q-value	PVE (%)
per se							
ASI-HN	S1_13685600	1	1.11×10^{-11}	7.58	0.18	6.89×10^{-7}	12.11
	S2_190189512	2	1.49×10^{-8}	−7.36	0.34	4.61×10^{-4}	30.18
Testcross							
PHT-LN	S1_39752558	1	5.15×10^{-7}	6.70	0.16	0.01	2.53
	S1_104874404	1	2.49×10^{-9}	8.80	0.39	1.54×10^{-4}	24.81
	S1_235704086	1	1.74×10^{-7}	−6.06	0.36	5.40×10^{-3}	3.65
PHT-HN	S3_104138066	3	1.30×10^{-8}	9.86	0.09	4.03×10^{-4}	4.46
	S3_179633217	3	5.75×10^{-7}	−8.49	0.13	0.01	19.19
	S6_165585769	6	2.04×10^{-9}	10.56	0.14	1.27×10^{-4}	1.28
GY-HN	S2_209927372	2	5.44×10^{-7}	−0.71	0.09	0.03	40.36

The q-value given is the chromosome-wide FDR-adjusted p-value.
PVE, phenotypic variance explained; GY, Grain yield (t ha^{-1}); PHT, plant height (cm); ASI, anthesis to silking interval (GDU).

observed a significant reduction in GY of BGEM lines and their derived testcrosses when evaluated under LN conditions, confirming the importance of sufficient N supply in maize production. Previous studies reported maize yield losses under N stress ranging from 37% to 78% (Bertin and Gallais, 2000; Presterl et al., 2002; Gallais and Hirel, 2003; Presterl et al., 2003; Abdel-Ghani et al., 2013; Chen et al., 2013; Das et al., 2019). In addition, testcross genotypes had better performance under LN than *per se* genotypes as a consequence of heterosis effect (Hallauer et al., 2010).

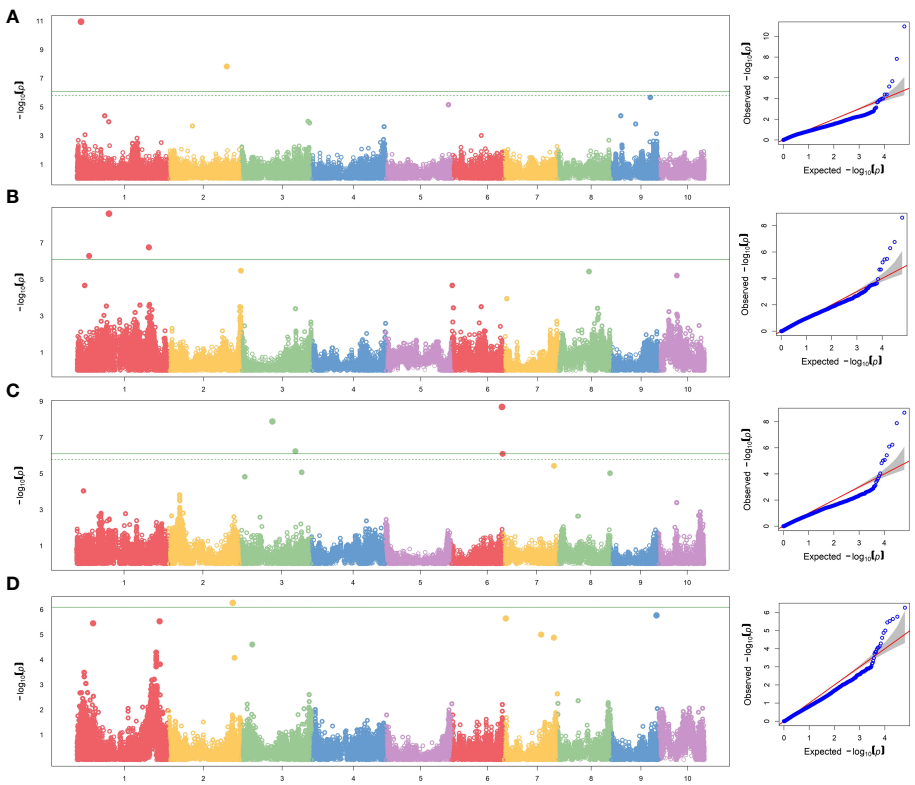


FIGURE 1
GWAS-derived Manhattan and QQ plots showing significant SNPs associated with **(A)** *per se* anthesis-silking interval under HN, **(B)** testcrosses plant height under LN and **(C)** under HN, and **(D)** testcrosses yield under HN using FarmCPU model. Each dot represents an SNP. The horizontal solid line represents the Bonferroni-corrected significant threshold of 8.10×10^{-7}.

10.3389/fpls.2023.1270166

TABLE 6 Candidate genes associated with agronomic traits of BGEM lines *per se*, and their testcrosses, grown under high nitrogen (HN) and low nitrogen (LN) conditions.

Traits	Chr	Gene start (bp)	Gene ID MaizeGDB	Gene ID Gramene	Gene name	Annotation
per se						
ASI - HN	1	13809656	Zm00001d027800	GRMZM2G169280	*ppr*	pentatricopeptide repeat-containing protein
	1	14081196	Zm00001d027808	GRMZM2G037912	*vap726*	putative vesicle-associated membrane protein 726
	1	13863300	Zm00001d027802	GRMZM2G004641	*hb64*	Homeobox-transcription factor 64
	2	190605384	Zm00001d005843	GRMZM2G088242	*hsftf2*	HSF-transcription factor 2
	2	190556326	Zm00001d005841	GRMZM2G414252	*bhlh20*	bHLH-transcription factor 20
	2	190962154	Zm00001d005856	GRMZM2G134502	*nup58*	nucleoporin58
Testcrosses						
PHT - LN	1	105862947	Zm00001d030103	GRMZM2G070271	*umc2230*	probable xyloglucan endotransglucosylase/hydrolase protein 27
	1	105553409	Zm00001d030098	GRMZM2G158976	*vq6*	VQ motif-transcription factor6
PHT - HN	3	177266069	Zm00001d042694	GRMZM2G110897	*poll1*	pollux-like1
	3	177609579	Zm00001d042706	GRMZM2G087619	*pds5a*	sister chromatid cohesion protein PDS5 homolog A
	3	177276266	Zm00001d042695	GRMZM2G110922	*snrkII4*	SnRK2 serine threonine protein kinase 4
	3	177338945	Zm00001d042697	GRMZM2G077333	*psbs1*	photosystem II subunit PsbS1
YLD - HN	2	209512508	Zm00001d006476	GRMZM2G171707	*aco5*	aconitase5
	2	209563817	Zm00001d006479	GRMZM2G168706	*cdpk3*	calcium dependent protein kinase3
	2	209688288	Zm00001d006486	GRMZM2G311187	*prh79*	protein phosphatase homolog79
	2	210172903	Zm00001d006508	GRMZM2G125495	*glr3.4*	glutamate receptor 3.4
	2	209822231	Zm00001d006493	GRMZM2G470075	*mate21*	multidrug and toxic compound extrusion21
	2	210284963	Zm00001d006512	GRMZM2G067063	*pdi12*	protein disulfide isomerase12

N deficiency is an important factor causing low yields in maize. During reproductive stage, N stress induces plant senescence, protein degradation, and thus reduces photosynthesis (Mu and Chen, 2021). To keep high GY in LN conditions, it is crucial to select genotypes with better performance under N stress conditions. Our study identified BGEM lines with outstanding performance under LN conditions. This shows the effectiveness of the DH technique in creating genetic variation that can be exploited in breeding for LN stress tolerance. Furthermore, the high performing lines from the same heterotic group could be used to develop breeding populations, either a synthetic population and/or several biparental populations. These could be used as a germplasm source for the development of new maize inbred lines with high allele frequency for NUE. Conversely, the BGEM lines from opposite heterotic groups might be used as parents in the development of maize hybrids tolerant to N stress conditions.

On average, the increase in ASI due to N deficiency stress was 16.24 GDUs in the per se trials and 20.01 GDUs in the testcross trials. Other studies have also reported an increase in ASI under LN conditions (Lafitte and Edmeades, 1995; Bertin and Gallais, 2000; Presterl et al., 2002; Gallais and Hirel, 2003; Abdel-Ghani et al., 2013; Das et al., 2019). According to Lin and Tsay (2017), the flowering time is postponed by either extreme deficiency or excess of N, while intermediate N concentrations promote flowering. Conversely, PHT means were lower under LN conditions for both *per se* and testcross trials. PHT reduction due to N deficiency stress was also observed in both inbred lines *per se* and testcrosses by Presterl et al. (2002). N is the most limiting nutrient and its rate of application influences maize growth and development at different stages. According to Singh et al. (2022), maize plants grown under LN conditions exhibited visual symptoms of N deficiency such as stunted growth and a significant reduction in shoot biomass. This indicates stress-related growth retardation, highlighting the prominent role of N for biomass accumulation (Qi and Pan, 2022).

Broad sense heritability in LN condition decreased from 0.02 (PHT in the combined analysis) to 0.52 (GY in Ames 2014) in the *per se* trials, and from 0.20 (PHT in the combined analysis) to 0.32 (GY in Ames 2015A) in the testcross trials. Decrease in heritability under stress conditions was also observed in previous studies in both maize inbred lines *per se* (Agrama et al., 1999; Bertin and Gallais, 2000; Gallais and Coque, 2005) and testcrosses (Bänziger et al., 1997; Presterl et al., 2002). Reasons for the decrease in heritability estimates include the decrease in genotypic variances instead of increased error variances (Bänziger et al., 1997; Gallais and Coque, 2005) and higher genotypes by environments interaction under LN than under HN (Gallais and Coque, 2005).

The significant genotype × N condition interactions for most of the traits suggests that the genotypes responded differently to the N conditions. According to Presterl et al. (2003), the high variance in the genotype × N interactions emphasizes the need for multi-environment testing to identify N-use efficient cultivars with a broad adaptation to different N levels.

4.2 Correlations between *per se* and testcross agronomic traits

Indirect selection for GY based on secondary traits is a cheaper approach compared to direct selection for GY due to relatively high heritability of secondary traits and high genetic correlation between secondary traits and GY under LN conditions. As heritability estimates for GY were low to moderate in our study, significant and close correlations between GY and traits with higher heritability, such as PHT and ASI, would be useful for indirect selection. Moreover, correlations between GY under HN and LN would be useful to predict GY under LN based on HN trials. However, the efficiency of indirect selection depends on the strength of the genetic correlation between the environments or traits. In this context, despite positive correlation between HN and LN conditions for GY, the magnitude of the correlation coefficients in our study was small and non-significant in most cases. While in the *per se* trials, the GY correlation between HN and LN was close to 0.70, the correlation was<0.50 in the testcross trials. This reveals how critical it is to evaluate genotypes under the target environment, for both stress and optimal N conditions. Indirect selection for GY under LN through performances obtained from HN conditions was found to be inefficient in a study conducted by Ertiro et al. (2020). According to the authors, low efficiency of indirect selection was explained by the low correlation between environments that resulted from a high proportion of genotype × N variance.

In our study, significant and moderately negative correlations were observed between GY and ASI in the *per se* trials, while these were not significant in the testcross trials. Silva et al. (2022) also reported a negative association between ASI and GY. Gallais and Hirel (2003) suggested that ASI may have a role in stress tolerance physiology, wherein having a shorter ASI would translate to that genotype having a better N metabolism efficiency, or increased yield under LN conditions. Correlations between PHT under HN and PHT under LN were higher than 0.70 for both *per se* and testcrosses trials, which indicates a possibility to evaluate PHT under only one N condition.

In terms of correlation between traits in BGEM lines *per se* and testcrosses, weak to non-significant correlations were observed between *per se* and testcross data. Therefore, the prediction of testcross performance based on *per se* information does not seem to be feasible for BGEM materials. This prediction is even more difficult for traits showing high heterotic effect, such as GY. Therefore, while the BGEM *per se* lines are mainly under additive genetic control, their testcrosses have the effect of dominance and, potentially, epistasis effects. According to Mihaljevic et al. (2005), an indirect improvement of testcross based on *per se* performance is economically advantageous, but it is only feasible with a high positive correlation between *per se* and testcross performance.

4.3 Significant SNP-trait associations detected by GWAS

The MLM model did not detect significant SNPs. The MLM with PCA and K model includes the kinship matrix in the model and is expected to reduce the false positives that arise from family relatedness (Yu et al., 2006). However, advantages of the MLM model to control false positives disappear for complex traits when they are associated with population structure having extensive genetic divergence. Kaler et al. (2019) reported that MLM model was particularly conservative and did not find any significant markers, while the FarmCPU model performed better with a less conservative approach. We used FarmCPU model, a GWAS approach that included population structure and kinship and additional algorithms that were used to address confounding problems between the markers and covariates Liu X. et al. (2016). This makes FarmCPU a GWAS approach that is intermediate between GLM and MLM in terms of stringency. In this context, the majority of candidate genes found in our study are related to stress tolerance. The *bHLH* (Zm00001d005841) displayed a subset of stress-responsive genes in *Arabidopsis* (Smolen et al., 2002). We also found a nuclear pore complex, *nup*, (Zm00001d005856), which is the main transport channel between cytoplasm and nucleoplasm and plays an important role in stress response. According to Liu et al. (2022), the overexpression of *nup58* in maize significantly promoted both chlorophyll content and activities antioxidant enzymes under drought and salt conditions. In addition, the expression patterns of the VQ genes (Zm00001d030098) have been analyzed in stress response in maize. According to Song et al. (2015), VQ motif-containing proteins play crucial roles in abiotic stress responses in plants. The expression profiles of VQ genes were analyzed in response to LN stress in soybean (Wang et al., 2014). The SnRK2 family members (Zm00001d042695) are plant-specific serine/threonine kinases involved in plant response to abiotic stresses and abscisic-acid-dependent plant development (Kulik et al., 2011). The *cdpk* (Zm00001d006479) is one of the well-known Ca^{2+} sensor protein kinases involved in environmental stress resistance (Asano et al., 2012). Several *cdpks* have been shown to be essential factors in abiotic stress tolerance, positively or negatively regulating stress tolerance by modulating abscisic acid signaling and reducing the accumulation of reactive oxygen species (Asano et al., 2012).

In our study, we found one SNP marker (S2_209927372) with over 40% of PEV. Although the literature reports few cases of total PEV higher than 30% for GY (Ajnone-Marsan et al., 1995; Sibov et al., 2003; Liu et al., 2012), the identification of a major GY-associated QTL is unusual. Fundamentally, a significant SNP can be due to a superior allele with potential to increase GY in elite germplasm. Conversely, a significant SNP can be caused by a yield-reducing allele. The latter option seems likely, given that GEM materials are based on non-adapted exotic introgressions. In addition, we observed that the SNPs found under LN did not overlap those found under HN. This result validates the low correlation observed between environments. Under abiotic stress conditions, the physiological mechanisms involved and genes responsible in control of traits may be different. Plants respond to

abiotic stress through a variety of physiological, biochemical, and transcriptional mechanisms Waters et al., 2017. Potentially, the genes exhibited altered levels of expression in response to the LN stress, which confirmed the need to screen and select genotypes for each N condition separately. We also observed negative and positive allelic effects. A positive value of allelic effect indicates that the minor allele was the favorable allele associated with the increase in the target trait, and a negative value indicates that the major allele was the favorable allele associated with the target trait (Ertiro et al., 2020).

Our derived DH lines may be promising materials for further studies on NUE or developing lines with improved NUE. SNPs significantly associated with agronomic traits under LN conditions, which can aid in improving NUE in maize. These SNPs can also be used to select for donor lines or superior breeding lines, after validating these putative SNPs by developing near-isogenic lines for linkage or expression analysis, or through transgenic methods. Our study shows that exotic germplasm from the GEM project are, therefore, useful sources of novel genes to select for yield and other agronomic traits under low N to improve NUE in maize.

Data availability statement

The original contributions presented in the study are publicly available. This data can be found here: https://doi.org/10.25380/iastate.24009039.v1.

Author contributions

DS: Data curation, Formal Analysis, Methodology, Writing – original draft, Writing – review & editing. AS: Formal Analysis, Writing – original draft, Writing – review & editing. PM: Writing – original draft, Writing – review & editing. EP: Methodology, Writing – original draft, Writing – review & editing. GD: Supervision, Validation, Writing – review & editing. MC: Funding acquisition, Validation, Writing – review & editing. MB: Funding acquisition, Writing – review & editing. TL: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The authors declare financial support was received for the research, authorship, and/or publication of this article. Funding for this work was provided by USDA's National Institute of Food and Agriculture (NIFA) Project, Nos. IOW04314, IOW01018, and IOW05510; and NIFA award 2018-51181-28419. Funding for this work was also provided by the R.F. Baker Center for Plant Breeding, Plant Sciences Institute, and K.J. Frey Chair in Agronomy at Iowa State University.

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

Abdel-Ghani, A. H., Kumar, B., Reyes-Matamoros, J., Gonzales-Portilla, P., Jansen, C., San Martin, J. P., et al. (2013). Genotypic variation and relationships between seedling and adult plant traits in maize (*Zea mays* L.) inbred lines grown under contrasting nitrogen levels. *Euphytica* 189, 123–133. doi: 10.1007/s10681-012-0759-0

Agrama, H. A., Zakaria, A. G., Said, F. B., and Tuinstra, M. (1999). Identification of quantitative trait loci for nitrogen use efficiency in maize. *Mol. Breed.* 5 (2), 187–195. doi: 10.1023/A:1009669507144

Ajnone-Marsan, P., Monfredini, G., Ludwig, W. F., Melchinger, A. E., Franceschini, P., Pagnotto, G., et al. (1995). In an elite cross of maize a major quantitative trait locus controls one-fourth of the genetic variation for grain yield. *Theoret. Appl. Genet.* 90, 415–424. doi: 10.1007/BF00221984

Alexander, R. B., Smith, R. A., Schwarz, G. E., Boyer, E. W., Nolan, J. V., and Brakebill, J. W. (2007). Differences in phosphorus and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin. *Environ. Sci. Technol.* 42 (3), 822–830. doi: 10.1021/es0716103

Andorf, C., Beavis, W. D., Hufford, M., Smith, S., Suza, W. P., Wang, K., et al. (2019). Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* 132, 817–849. doi: 10.1007/s00122-019-03306-3

Asano, T., Hayashi, N., Kikuchi, S., and Ohsugi, R. (2012). CDPK-mediated abiotic stress signaling. *Plant Signal. Behav.* 7 (7), 817–821. doi: 10.4161/psb.20351

Bänziger, M., Betrán, F. J., and Lafitte, H. R. (1997). Efficiency of high-nitrogen selection environments for improving maize for low-nitrogen target environments. *Crop Sci.* 37, 1103–1109. doi: 10.2135/cropsci1997.0011183X003700040012x

Bertin, P., and Gallais, A. (2000). Genetic variation for nitrogen use efficiency in a set of recombinant maize inbred lines. I. Agrophysiological results. *Maydica* 45. doi: 10.3389/fpls.2021.625915

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. *Bioinform* 23 (19), 2633–2635. doi: 10.1093/bioinformatics/btm308

Brenner, E. A., Blanco, M., Gardner, C., and Lübberstedt, T. (2012). Genotypic and phenotypic characterization of isogenic doubled haploid exotic introgression lines in maize. *Mol. Breed.* 30, 1001–1016. doi: 10.1007/s11032-011-9684-5

Chen, F. J., Fang, Z. G., Gao, Q., Ye, Y. L., Jia, L. L., Yuan, L. X., et al. (2013). Evaluation of the yield and nitrogen use efficiency of the dominant maize hybrids grown in North and Northeast China. *Sci. China Life Sci.* 56, 552–560. doi: 10.1007/s11427-013-4462-8

Das, B., Atlin, G. N., Olsen, M., Burgueño, J., Tarekegne, A., Babu, R., et al. (2019). Identification of donors for low-nitrogen stress with maize lethal necrosis (MLN) tolerance for maize breeding in sub-Saharan Africa. *Euphytica* 215, 80. doi: 10.1007/s10681-019-2406-5

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloSOne* 6 (5), e19379. doi: 10.1371/journal.pone.0019379

Ertiro, B. T., Olsen, M., Das, B., Gowda, M., and Labuschagne, M. (2020). Efficiency of indirect selection for grain yield in maize (*Zea mays* L.) under low nitrogen conditions through secondary traits under low nitrogen and grain yield under optimum conditions. *Euphytica* 216, 134. doi: 10.1007/s10681-020-02668-w

Gallais, A., and Coque, M. (2005). Genetic variation and selection for nitrogen use efficiency in maize: a synthesis. *Maydica* 50, 531–547.

Gallais, A., and Hirel, B. (2003). An approach to the genetics of nitrogen use efficiency in maize. *J. Exp. Bot.* 55, 295–306. doi: 10.1093/jxb/erh006

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., and Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genet. Epidemiol.* 34, 100–105. doi: 10.1002/gepi.20430

Getahun, B. B., Visser, R. G. F., and van der Linden, C. G. (2020). Identification of QTLs associated with nitrogen use efficiency and related traits in a diploid potato population. *Am. J. Potato Res.* 97, 185–201. doi: 10.1007/s12230-020-09766-4

Gheith, E. M. S., El-Badry, O. Z., Lamlom, S. F., Ali, H. M., Siddiqui, M. H., Ghareeb, R. Y., et al. (2022). Maize (*Zea mays* L.) productivity and nitrogen use efficiency in response to nitrogen application levels and time. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.941343

Goolsby, D. A., Battaglin, W. A., Aulenbach, B. T., and Hooper, R. P. (2000). Nitrogen flux and sources in the Mississippi River Basin. *Sci. Total Environ.* 248 (2), 75–86. doi: 10.1016/s0048-9697(99)00532-x

Hallauer, A. R., Miranda Filho, J. B., and Carena, M. J. (2010). *Quantitative genetics in maize breeding* (New York: Springer).

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., et al. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genom.* 11, 724. doi: 10.1186/1471-2164-11-724

Kaler, A. S., Gillman, J. D., Beissinger, T., and Purcell, L. C. (2019). 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

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030. doi: 10.1093/database/bar030

Kulik, A., Wawer, I., Krzywińska, E., Bucholc, M., and Dobrowolska, G. (2011). SnRK2 protein kinases - key regulators of plant response to abiotic stresses. *Omics J. Integr. Biol.* 15 (12), 859–872. doi: 10.1089/omi.2011.0091

Kumari, S., Sharma, N., and Raghuram, N. (2021). Meta-analysis of yield-related and n-responsive genes reveals chromosomal hotspots, key processes and candidate genes for nitrogen-use efficiency in rice. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.627955

Ladha, J. K., Tirol-Padre, A., Reddy, C. K., Cassman, K. G., Verma, S., Powlson, D. S., et al. (2016). Global nitrogen budgets in cereals: a 50-year assessment for maize, rice, and wheat production systems. *Sci. Rep.* 6, 19355. doi: 10.1038/srep19355

Lafitte, H. R., and Edmeades, G. O. (1995). Association between traits in tropical maize inbred lines and their hybrids under high and low soil nitrogen. *Maydica* 40 (3), 259–267.

Lin, Y. L., and Tsay, Y. F. (2017). Influence of differing nitrate and nitrogen availability on flowering control in Arabidopsis. *J. Exp. Bot.* 68, 2603–2609. doi: 10.1093/jxb/erx053

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

Liu, Z., Abou-Elwafa, S. F., Xie, J., Liu, Y., Li, S., Aljabri, M., et al. (2022). A Nucleoporin NUP58 modulates responses to drought and salt stress in maize (*Zea mays* L.). *Plant Sci.* 320, 111296. doi: 10.1016/j.plantsci.2022.111296

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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Liu, R., Jia, H., Cao, X., Huang, J., Li, F., Tao, Y., et al. (2012). Fine mapping and candidate gene prediction of a pleiotropic quantitative trait locus for yield-related trait in *Zea mays*. *PloS One* 7 (11), e49836. doi: 10.1371/journal.pone.0049836

Liu, Z., Wang, Y., Ren, J., Mei, M., Frei, U. K. B., Trampe, B., et al. (2016a). Maize doubled haploids. *Plant Breed. Rev.* 40, 123. doi: 10.1002/9781119279723.ch3

Liu, Z., Zhu, C., Jiang, Y., Tian, Y., Yu, J., An, H., et al. (2016b). Association mapping and genetic dissection of nitrogen use efficiency-related traits in rice (*Oryza sativa* L.). *Funct. Integr. Genomics* 16, 323–333. doi: 10.1007/s10142-016-0486-z

Ma, L., Qing, C., Frei, U., Shen, Y., and Lübberstedt, T. (2020). Association mapping for root system architecture traits under two nitrogen conditions in germplasm enhancement of maize doubled haploid lines. *Crop J.* 8 (2), 213–226. doi: 10.1016/j.cj.2019.11.004

Ma, L., Wang, C., Hu, Y., Dai, W., Liang, Z., Zou, C., et al. (2022). GWAS and transcriptome analysis reveal MADS26 involved in seed germination ability in maize. *Theor. Appl. Genet.* 135 (5), 1717–1730. doi: 10.1007/s00122-022-04065-4

McGeehan, S. L., and Naylor, D. V. (1988). Automated instrumental analysis of carbon and nitrogen in plant and soil samples. *Commun. Soil Sci. Plant Anal.* 19 (4), 493–505. doi: 10.1080/00103628809367953

Mihaljevic, R., Schön, C. C., Utz, H. F., and Melchinger, A. E. (2005). Correlations and QTL correspondence between line per se and testcross performance for agronomic traits in four populations of European maize. *Crop Sci.* 45, 114–122. doi: 10.2135/cropsci2005.0114a

Mu, X., and Chen, Y. (2021). The physiological response of photosynthesis to nitrogen deficiency. *Plant Physiol. Biochem.* 158, 76–82. doi: 10.1016/j.plaphy.2020.11.019

Nag, P., and Das, S. (2022). "Microbiome to the rescue: nitrogen cycling and fixation in non-legumes," in *Nitrogen fixing bacteria: sustainable growth of non-legumes* (Singapore: Springer Nature Singapore), 195–214.

Pollak, L. M. (2003). The history and success of the public–private project on germplasm enhancement of maize (GEM). *Adv. Agron.* 78, 45–87. doi: 10.1016/S0065-2113(02)78002-4

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* 47, 1146–1154. doi: 10.1093/nar/gky1046

Presterl, T., Seitz, G., Landbeck, M., Thiemt, E. M., Scmidt, W., and Geiger, H. H. (2003). Improving nitrogen-use efficiency in European maize - estimation of quantitative genetic parameters. *Crop Sci.* 43, 1259–1265. doi: 10.2135/cropsci2003.1259

Presterl, T., Seitz, G., Scmidt, W., and Geiger, H. H. (2002). Improving nitrogen-use efficiency in European maize – comparison between line *per se* and testcross performance under high and low soil nitrogen. *Maydica* 47, 83–91.

Qi, D., and Pan, C. (2022). Responses of shoot biomass accumulation, distribution, and nitrogen use efficiency of maize to nitrogen application rates under waterlogging. *Agric. Water Manage.* 261, e107352. doi: 10.1016/j.agwat.2021.107352

R Core Team (2014). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing).

Röber, F. K., Gordillo, G. A., and Geiger, H. H. (2005). *In vivo* haploid induction in maize-performance of new inducers and significance of doubled haploid lines in hybrid breeding. *Maydica* 50, 275–283.

Salhuana, W., and Pollak, L. (2006). Latin American maize project (LAMP) and germplasm enhancement of maize (GEM) project: Generating useful breeding germplasm. *Maydica* 51 (2), 339–355.

Sanchez, D. L., Liu, S., Ibrahim, R., Blanco, M., and Lübberstedt, T. (2018). Genome-wide association studies of doubled haploid exotic introgression lines for root system architecture traits in maize (*Zea mays* L.). *Plant Sci.* 268, 30–38. doi: 10.1016/j.plantsci.2017.12.004

Sibov, S. T., Souza, C. L., Garcia, A. A. F., Silva, A. R., Garcia, A. F., Mangolin, C. A., et al. (2003). Molecular mapping in tropical maize (*Zea mays* L.) using microsatellite markers. *Hereditas* 139 (2), 107–115. doi: 10.1111/j.1601-5223.2003.01667.x

Silva, P. C., Sánchez, A. C., Opazo, M. A., Mardones, L. A., and Acevedo, E. A. (2022). Grain yield, anthesis-silking interval, and phenotypic plasticity in response to changing environments: Evaluation in temperate maize hybrids. *Field Crops Res.* 285, e108583. doi: 10.1016/j.fcr.2022.108583

Singh, P., Kumar, K., Jha, A. K., Yadava, P., Pal, M., Rakshit, S., et al. (2022). Global gene expression profiling under nitrogen stress identifies key genes involved in nitrogen stress adaptation in maize (*Zea mays* L.). *Sci. Rep.* 12, 4211. doi: 10.1038/s41598-022-07709-z

Smolen, G. A., Pawlowski, L., Wilensky, S. E., and Bender, J. (2002). Dominant alleles of the basic helix-loop-helix transcription factor ATR2 activate stress-responsive genes in Arabidopsis. *Genetics* 161 (3), 1235–1246. doi: 10.1093/genetics/161.3.1235

Song, W., Zhao, H., Zhang, X., Lei, L., and Lai, J. (2015). Genome-wide identification of VQ motif-containing proteins and their expression profiles under abiotic stresses in maize. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.01177

Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Lipka, A. E., Flint-Garcia, S., et al. (2018). Association mapping of flowering and height traits in germplasm enhancement of maize doubled haploid (GEM-DH) lines. *Plant Genome* 11 (2), 170083. doi: 10.3835/plantgenome2017.09.0083

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 (2), 147–157. doi: 10.1002/cppb.20050

Verzegnazzi, A. L., Dos Santos, I. G., Krause, M. D., Hufford, M., Frei, U. K., Campbell, J., et al. (2021). Major locus for spontaneous haploid genome doubling detected by a case–control GWAS in exotic maize germplasm. *Theor. Appl. Genet.* 134, 1423–1434. doi: 10.1007/s00122-021-03780-8

Wang, X., Zhang, H., Sun, G., Jin, Y., and Qiu, L. (2014). Identification of active VQ motif-containing genes and the expression patterns under low nitrogen treatment in soybean. *Gene* 543, 237–243. doi: 10.1016/j.gene.2014.04.012

Wani, S. H., Vijayan, R., Choudhary, M., Kumar, A., Zaid, A., Singh, V., et al. (2021). Nitrogen use efficiency (NUE): elucidated mechanisms, mapped genes and gene networks in maize (*Zea mays* L.). *Physiol. Mol. Biol. Plants* 27, 2875–2891. doi: 10.1007/s12298-021-01113-z

Waters, A. J., Makarevitch, I., Noshay, J., Burghardt, L. T., Hirsch, C. N., Hirsch, C. D., et al. (2017). Natural variation for gene expression responses to abiotic stress in maize. *Plant J.* 89 (4), 706–717. doi: 10.1111/tpj.13414

Wu, L., Zheng, Y., Jiao, F., Wang, M., Zhang, J., Zhang, Z., et al. (2022). Identification of quantitative trait loci for related traits of stalk lodging resistance using genome-wide association studies in maize (*Zea mays* L.). *BMC Genom. Data* 23, 1–16. doi: 10.1186/s12863-022-01091-5

Xu, S., Tang, X., Zhang, X., Wang, H., Ji, W., Xu, C., et al. (2023). Genome-wide association study identifies novel candidate loci or genes affecting stalk strength in maize. *Crop J.* 11, 220–227. doi: 10.1016/j.cj.2022.04.016

Yu, X., Keitel, C., Zhang, Y., Wangeci, A. N., and Dijkstra, F. A. (2022). Global meta-analysis of nitrogen fertilizer use efficiency in rice, wheat and maize. *Agric. Ecosyst. Environ.* 338, e108089. doi: 10.1016/j.agee.2022.108089

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., 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), 203–2088. doi: 10.1038/ng1702

Zuffo, L. T., DeLima, R. O., and Lübberstedt, T.. (2022). Combining datasets for maize root seedling traits increases the power of GWAS and genomic prediction accuracies. *J. Exp. Bot.* 73 (16), 5460–5473. doi: 10.1093/jxb/erac236

frontiers | Frontiers in Plant Science

TYPE Methods
PUBLISHED 10 October 2023
DOI 10.3389/fpls.2023.1221790

OPEN ACCESS

EDITED BY
Hakan Ozkan,
Çukurova University, Türkiye

REVIEWED BY
Kaushik Ghose,
Texas Tech University, United States
Peng Cheng,
Zhejiang Academy of Agricultural Sciences, China
Usman Aziz,
Northwestern Polytechnical University, China

*CORRESPONDENCE
Hai Thanh Tran
haitranthanhclrri@gmail.com
Peter A. Anderson
peter.anderson@flinders.edu.au

RECEIVED 12 May 2023
ACCEPTED 22 September 2023
PUBLISHED 10 October 2023

CITATION
Tran HT, Schramm C, Huynh M-m, Shavrukov Y, Stangoulis JCR, Jenkins CLD and Anderson PA (2023) An accurate, reliable, and universal qPCR method to identify homozygous single insert T-DNA with the example of transgenic rice. *Front. Plant Sci.* 14:1221790. doi: 10.3389/fpls.2023.1221790

An accurate, reliable, and universal qPCR method to identify homozygous single insert T-DNA with the example of transgenic rice

Hai Thanh Tran*, Carly Schramm, My-my Huynh, Yuri Shavrukov, James C. R. Stangoulis, Colin L. D. Jenkins and Peter A. Anderson*

College of Science and Engineering, Flinders University, Adelaide, SA, Australia

Early determination of transgenic plants that are homozygous for a single locus T-DNA insert is highly desirable in most fundamental and applied transgenic research. This study aimed to build on an accurate, rapid, and reliable quantitative real-time PCR (qPCR) method to fast-track the development of multiple homozygous transgenic rice lines in the T_1 generation, with low copy number to single T-DNA insert for further analyses. Here, a well-established qPCR protocol, based on the *OsSBE4* reference gene and the *nos* terminator, was optimized in the transgenic Japonica rice cultivar Nipponbare, to distinguish homozygous single-insert plants with 100% accuracy. This method was successfully adapted to transgenic Indica rice plants carrying three different T-DNAs, without any modifications to quickly develop homozygous rice plants in the T_1 generation. The accuracy of this qPCR method when applied to transgenic Indica rice approached 100% in 12 putative transgenic lines. Moreover, this protocol also successfully detected homozygous single-locus T-DNA transgenic rice plants with two-transgene T-DNAs, a feature likely to become more popular in future transgenic research. The assay was developed utilizing universal primers targeting common sequence elements of gene cassettes (the *nos* terminator). This assay could therefore be applied to other transgenic plants carrying the *nos* terminator. All procedures described here use standardized qPCR reaction conditions and relatively inexpensive dyes, such as SYBR Green, thus the qPCR method could be cost-effective and suitable for lower budget laboratories that are involved in rice transgenic research.

KEYWORDS

quantitative real-time PCR, zygosity identification, homozygous plant, *NOS* terminator, *OsSBE4* reference gene, single T-DNA, copy number, transgenic rice

Introduction

An efficient and productive protocol to identify homozygous plants plays a crucial role in molecular genetics and physiological studies, delivery of CRISPR-Cas9 cassettes for genome editing, and for molecular breeding of transgenic plants such as rice. A successful transformation process produces a first generation (T_0) of hemizygous plants, where the T-DNA is inserted without an allelic counterpart. The T_0 plants may contain one, two or multiple copies of the T-DNA in the host genome at the same or different locations. After self-pollination, the T-DNA segregates according to Mendelian principles. This means that the progeny in the subsequent generation (T_1) are either homozygous, hemizygous or null for a particular T-DNA insert. To ensure stable inheritance of the T-DNA and any ongoing genetic or physiological analysis, only homozygous plants should be used, and preferably those with a single T-DNA insert. Thus, development of homozygous transgenic lines is a fundamental requirement in most of the downstream applications and studies of transgenic organisms, including plants. Therefore, an efficient and rapid technique for determining the T-DNA copy number, and screening for homozygous plants in the T_1 generation, would fast-track this selection process.

To date, Southern blot and quantitative real-time PCR (qPCR) are the two most common methods for quantifying T-DNA copy number in transgenic plants. Southern blot analysis (Southern, 1975) is a powerful and reliable method but requires a large amount of genomic DNA and is time-consuming, in some cases hazardous and laborious, and thus is expensive. Especially for rice and other cereal crops which have a long-life cycle (4-5 months), large amount of space and resources are required for growth to generate subsequent generations. An alternatively to Southern blot analysis is qPCR, a rapid technique that is based on the detection of fluorescence generated during the amplification process (German et al., 2003; Bubner and Baldwin, 2004; Bubner et al., 2004). The fluorescence can be produced by an intercalating dye that fluoresces as it binds to double-stranded DNA or using a probe containing both a fluorophore and a quencher (TaqMan) targeted to an internal region of the transgene. During the extension step in the latter qPCR assay, DNA polymerase degrades the probe, releasing the fluorophore from the quencher and giving rise to fluorescence. The result of the PCR reaction is expressed as a cycle threshold (Ct) value. Other more technically sophisticated techniques have been developed from the basic TaqMan concept, but TaqMan requires relatively expensive probes (Baric et al., 2006) and it is less suitable for large-scale and universal testing applications. Droplet digital PCR (ddPCR) was evaluated and used to estimate T-DNA copy number in several crop species (Collier et al., 2017; Giraldo et al., 2019; Cai et al., 2020; Cai et al., 2021), homozygous transgenic tobacco (Głowacka et al., 2016), and maize (Xu et al., 2016). Again, ddPCR replies on an initial restriction digestion step, expensive probes and more expensive droplet digital PCR systems to estimate T-DNA copy number, and this thus is less suited to lower budget laboratories.

An alternative qPCR-based method known as the comparative Ct ($2^{-\Delta\Delta Ct}$) method is reported for determining copy number and/or zygosity of transgenes (Bubner and Baldwin, 2004). Many analyses on copy number and/or zygosity of transgenes have employed the qPCR technique in several species such as maize (Schmidt and Parrott, 2001; Song et al., 2002; Shou et al., 2004; Xu et al., 2016), tomato (Mason et al., 2002; German et al., 2003), tobacco (Bubner et al., 2004), wheat (Li et al., 2004; Gadaleta et al., 2011; Giancaspro et al., 2017), and rice (Yang et al., 2005). Most studies showed a robust effectiveness of copy number determination, but effective outcomes for zygosity were unclear, except for the studies of Ingham et al. (2001). Later, it was shown that qPCR could be used to screen for homozygous transgenic cereal crops (Mieog et al., 2013; Wang et al., 2015) but the previous protocols were not universally adaptable for different T-DNAs and they were not consistent in all plants with identical zygosity. In most transgenic plant breeding studies, a large number of transgenic lines transformed with different gene expression cassettes can be generated. Therefore, a universal and fast-tracking method of homozygous plant identification would be advantageous. Here, the universal qPCR assay was optimized for accurate and reliable determination of zygosity using the example of transgenic rice plants. Compared to previous methods, this standardized qPCR assay is cost-effective and useful as a high-throughput method for fast-tracking identification of homozygous plants carrying single or low insert numbers in a single generation. It was successfully adopted to identify many homozygous single- or two-insert T-DNA rice plants in the T_1 generation carrying four different gene expression cassettes, all with the same *nos* terminator sequence, without any modification.

Materials and methods

Agrobacterium-mediated transformation and transgenic plants

Multiple Japonica and Indica transgenic rice plants were generated via *Agrobacterium*-mediated transformation with four different T-DNAs (Figure 1). The transgenic Japonica lines carrying the *Act-1:HvSUT1:NosT* and *Glb-1:HvSUT1:NosT* cassettes were abbreviated to A lines and G lines, respectively. The Indica rice lines containing the *Glb-1:HvSUT1:NosT*, *GluA2:OsNAS2:NosT*, and *GluA2:OsNAS2:NosT-Glb-1:HvSUT1:NosT* were designated as SC1, SC2 and DC1, respectively. All T-DNAs contained a single common sequence of the nopaline synthase (*nos*) terminator, except for the DC1 T-DNA which contained two *nos* terminators. Illustrations of the T-DNAs were prepared using SnapGene (GSL Biotech).

Primer design

The rice genes coding for branching enzyme (*OsSBE4*) and sucrose phosphate synthase (*OsSPS*), are appropriate reference genes in the qPCR assay as they are each present as two copies at a single locus in the rice genome (Mizuno et al., 2001; Jiang et al., 2009). All oligonucleotide primers were designed using the Primer3 software (bioinfo.ut.ee/primer3-0.4.0/primer3/). Some parameters

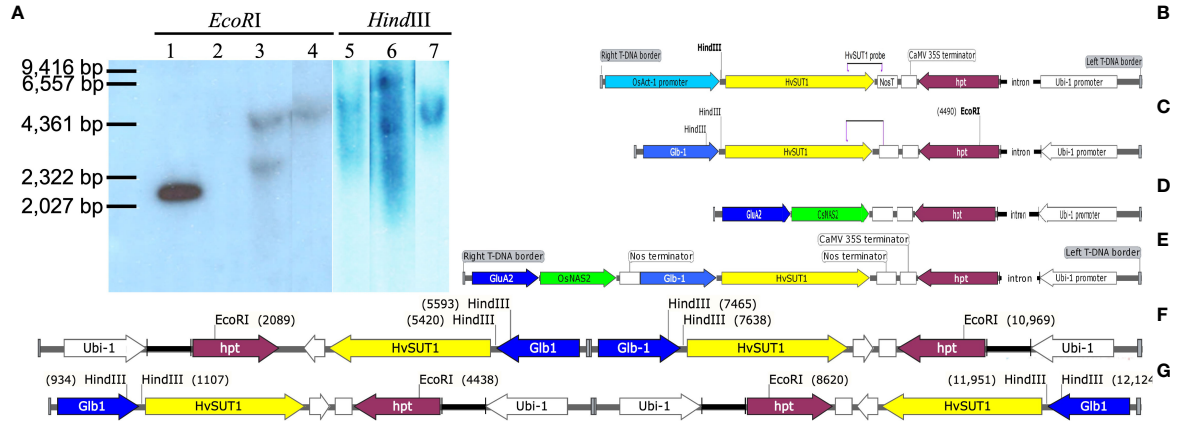


FIGURE 1
Southern blot analyses of the transgenic lines and schematic representation of the four T-DNAs used for rice transformation. **(A)** Genomic DNA of the T_0 transgenic lines was digested with *Hind*III or *Eco*RI, then probed with a fragment of the *HvSUT1* gene. Migration of *Hind*III-digested lambda marker is noted. Lane 1 contains 100 pg of the pIPKb001 plasmid (EU161567) carrying *Glb : HvSUT1:NosT* used as a positive control. Lane 2 is the genomic DNA of non-transgenic rice Nipponbare digested with *Eco*RI used as a negative control. Lane 3 and 5 showed two bands in the genomic DNA of the G1.4 transgenic line carrying the *Glb : HvSUT1:NosT* digested by *Eco*RI and a single band by *Hind*III, respectively. Lane 4 and 6 showed a single band in the genomic DNA of the G3.1 transgenic line carrying the *Glb : HvSUT1:NosT* digested by *Eco*RI and two bands by *Hind*III, respectively. Lane 7 showed a single band in the DNA genomic of the A5.1 transgenic line carrying the *Act-1:HvSUT1:NosT* digested by *Hind*III. The *Act-1:HvSUT1:NosT* construct with one *nos* terminator **(B)** and the T-DNA of the *Glb-1:HvSUT1:NosT* with one *nos* terminator **(C)** with the *Hind*III and *Eco*RI sites and a position of the probe of *HvSUT1*. The T-DNA containing two stacked genes with two *nos* terminator regions (the *Glb-1:HvSUT1:NosT* and *GluA2:OsNAS2:NosT)* **(D)** and the T-DNA of *GluA2:OsNAS2:NosT* with one *nos* terminator **(E)**. The T-DNA of the G3.1 transgenic line **(F)** and the G1.4 transgenic line **(G)** with the insertion of two T-DNAs in a tandem repeat, but inverted orientation. *Glb-1* and *GluA2*: rice endosperm specific promoter *Globulin 1* and *Glutelin A2*; *Act-1*: the rice actin 1 gene promoter; *HvSUT1*: barley sucrose transporter 1, *OsNAS2*: rice nicotianamine synthetase 2; hygromycin phosphotransferase (*hpt*) gene under control of the *Ubi-1* promoter and the *CaMV 35S* terminator (35S T). Illustrations of the DNAs were prepared using SnapGene (GSL Biotech).

were modified from the default setting to achieve similar amplification efficiencies between the reference genes and T-DNAs in the qPCR assay, including product size (80-200 bp) and primer Tm (59-65 °C). The selected primer pairs were checked by BLASTn against representative genomes for *Oryza sativa* ssp. Japonica and Indica to check the specificity of each primer pair. Folding template of the reference genes and T-DNAs were analysed by OligoArchitect™ Primer (offered by Sigma-Merck, USA) to eliminate primers that matched with predicted secondary structures in the templates. UNAFold software (http://sg.idtdna.com/UNAFold) was used to predict whether amplicons generated any secondary structure at the annealing temperature. Primer pairs that passed all of these requirements were synthesized by Sigma-Merck (Supplementary Table S1).

End-point PCR for confirmation of T-DNA integration

Rice genomic DNA (gDNA) used for end-point PCR and the qPCR assay was isolated from leaves (100 mg) of transgenic lines (T_0, T_1 or T_2) using an Isolate II plant DNA kit (Bioline, USA). The gDNA was suspended in 50 µl of elution buffer (5 mM Tris-HCl, pH 8.5) and the DNA samples were diluted to achieve an average DNA concentration of around 4.26 ng/ul using sterilized MiliQ water. The quantity of gDNA was estimated using a Nanodrop 2000 instrument (Thermo Fisher Scientific, USA). For end-point PCR (Supplementary Table S2), a 25 µl reaction consisted of 1 × GoTaq Green Reaction Buffer (Promega, USA), 2 mM $MgCl_2$, 0.4 mM dNTPs, 0.4 µM for each primer and 2 µl of 4.26 ng/µl DNA template. The PCR cycling protocol was as follows; initial denaturation for 10 min at 95 °C, followed by 30 cycles of 95 °C for 30 sec, 64 °C for 30 sec and 72 °C for 30 sec, ending with 10 min at 72 °C. PCR products were visualized on a 1.5% agarose gel.

Standard curve establishment

An efficient, reproducible, and dynamic range qPCR assay was determined by making a standard curve using a 10-fold serial dilution of the target sequence (10^4, 10^3, 10^2, 10^1 and 10^0 copies/µl). The absolute quantification of the copy number of the target sequence per haploid genome was calculated based on the mass of genomic DNA and the genomic DNA concentration determined by UV absorption at 260 nm based on the protocol of Applied Biosystems (Applied Biosystems, 2003). The mass of a single copy of the rice genome (haploid) was calculated as follows:

$$m = n \times (1.096 \times 10^{-21}) \text{ g/bp},$$

where m is mass (g) of genomic DNA and n is genome size (389 Mb) (Sequencing Project International Rice, 2005).

Then mass of the rice genomic DNA needed to achieve 10^4 copies of the rice genome was calculated as follows

$$mass\ of\ gDNA\ needed = m \times 10^4 \text{ (g)}.$$

For this case, 4.26 ng/µl was equivalent to 10^4 copies/µl of the rice genome. The stock concentration of rice gDNA was determined by UV absorption at 260 nm and diluted to achieve 4.26 ng/µl (10^4

copies/µl). The 10^4 copies/µl was serially diluted 10-fold in sterile Milli Q water to reach 10^3, 10^2, 10^1 and 10^0 copies/µl of the rice gDNA.

The standard curve was generated from Ct values plotted against the logarithm of the template concentration at each dilution with three replicates. Amplification efficiency was calculated from the slope of the standard curve (Bio-Rad Laboratories, 2006). The amplification efficiencies of the primer pairs for the reference genes and the T-DNAs were then compared with each other and chosen when within 5% of each other, and as close to 100% as possible. Based on the standard deviation of five dilutions in the standard curve, a concentration of DNA template was chosen with the lowest standard deviation and then used in the qPCR assay for zygosity analysis.

qPCR assay for zygosity identification

Reactions were done in a CFX96™ Real-time PCR detection instrument (BioRad, USA). A 10 µl reaction consisted of 5 µl (1 ×) KiCqStart SYBR Green qPCR ReadyMix (Sigma-Merck, USA), 300 nM for each primer and 2.5 µl of gDNA template (0.426 ng/µl ~10^3 copies/µl). A standard two-step protocol was used as follows: 1 cycle for DNA denaturation at 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 sec and 60 °C for 30 sec and a melting curve was generated in 0.5 °C increments starting at 60 °C. Reactions for both reference gene (*OsSBE4*) and the T-DNAs were run in triplicate.

As the amplification efficiency of the reference gene and T-DNAs were similar and nearly 100% and within 5% of each other, the zygosity of transgenic plants was determined by the Livak method, $2^{-\Delta\Delta Ct}$ (Livak and Schmittgen, 2001). The difference between the Ct values of the T-DNAs and reference gene for the test plant was calculated (ΔCt), then normalized to the ΔCt for a calibrator plant to obtain copy number calculation (ΔΔCt). The resulting ΔΔCt value was incorporated to determine zygosity of the test plant. The $2^{-\Delta\Delta Ct}$ method was calculated using the following steps.

First, the Ct values of the T-DNAs (for example *HvSUT1* and *Nos* terminator) were normalized to that of the reference gene (*OsSBE4*) for both test plants and calibrator:

$$\Delta Ct\ \ (test) = Ct\ \ (target,\ \ test) - Ct\ \ (ref, test)$$

ΔCt (calibrator)= Ct (target, calibrator)-Ct (ref, calibrator) .

Second, the ΔCt of test plant was normalized to that of the calibrator to obtain ΔΔCt:

$$\Delta\Delta Ct = \Delta Ct\ \ (test) - \Delta Ct\ \ (calibrator)\ .$$

Finally, the 2-fold difference was then found by 2 to the power of -ΔΔCt:

$$2^{-\Delta\Delta Ct}$$

If the calibrator was hemizygous (that is a single-insert T_0 plant), the $2^{-\Delta\Delta Ct}$ of a homozygous single-insert plant will be twice that of the calibrator.

Southern blot for determining insert number of T-DNAs

Southern blot analysis was conducted to provide a confirmation of T-DNA insert number independent to the qPCR assay. Rice gDNA of T_0 transgenic Japonica lines was digested with *Hind*III or *Eco*RI (New England BioLabs, USA) in 30 µl reactions which consisted of 3 µl buffer (10 ×), restriction enzyme (30 units) and 3 µg gDNA. The reaction was incubated at 37 °C for at least 4 hrs. Digested fragments were separated on a 0.8% agarose gel by electrophoresis at 30-40 V overnight. After the separation of the DNA fragments was completed, the DNA was depurinated by incubating the gel in 0.25 M HCl for 10 min. The gel was rinsed in MilliQ water three times before the DNA was denatured in the high salt denaturation solution for 30 min. Afterward the DNA was transferred directly to Biodyne B nylon membrane (Pall Life Sciences, USA) by capillary transfer overnight in 20 × SSC buffer (pH 7) containing sodium chloride (3 M) and sodium citrate buffer (0.3 M). The membrane was then dried at 80 °C for 15 min, and then fixed with UV cross-linker (GS Gene linker UV chamber, BioRad, USA) at 150 mJoule for 10 sec. After fixation, the membrane was processed with hybridization and detection.

A 498-bp fragment amplified from the *HvSUT1* plasmid was used as a probe template for radioactive ^{32}P-dCTP labelling using random primers (Geneworks, Australia), which consisted of 3 µl of random primer (40 µM), 5 µl of 20 ng/µl probe template, 12.5 µl of 2 × oligo labelling buffer, 1.5 µl of DNA polymerase (Klenow fragment) (2 units/µl) and 5 µl of ^{32}P-dCTP in a 27 µl reaction. The labelled probe was incubated at 37 °C for 1-2 hrs, and then isolated by ethanol precipitation. Before hybridization, the membrane was mixed well and incubated with 5 × SSC overnight in the hybridization oven at 65 °C. The hybridization and detection processes were carried out according to Weiss (1992).

Results

A well-optimized qPCR protocol for homozygous single-insert determination

A transgenic Japonica rice line overexpressing the *Act-1:HvSUT1:NosT* (designated A5.1) was confirmed to carry a single insert in its genome by Southern blot, following digestion of genomic DNA with *Hind*III, which cuts once in the T-DNA construct (Figures 1A, B). The genomic DNA of this single copy line was used to optimize conditions of the qPCR protocol. For the expression construct *Glb-1:HvSUT1:NosT* (transformed plants designated G lines), there was a single *Eco*RI recognition site in hygromycin phosphotransferase (*hpt*) gene near on the left T-DNA border and two *Hind*III recognition sites close to the *HvSUT1* gene near the right border of the T-DNA construct. Two Southern blot analyses of the G lines were carried out with *Hind*III and *Eco*RI digested gDNA. The G1.4 line showed two bands in *Eco*RI digested DNA and a single band in *Hind*III digested DNA, while the G3.1 line had two hybridising fragments in *Eco*RI digested DNA and one

fragment in the *Hind*III digestion (Figure 1A). The two G3.1 and G1.4 transgenic lines were predicted to contain two inserts of the T-DNA in a tandem repeat, but inverted orientation (Figures 1F, G).

To reliably determine the zygosity of transgenic lines using the qPCR protocol, the comparative $2^{-\Delta\Delta Ct}$ method was used for copy number estimation (Bubner and Baldwin, 2004). This method requires similar PCR efficiencies of the T-DNA amplicons and an endogenous reference gene present as two copies at a single locus in the diploid rice genome. Thus, specific and well-matched sets of primer pairs for the T-DNAs and reference genes were designed and it was verified that their amplification efficiencies were close to 100% ± 5%. Three primer pairs for two endogenous reference genes, namely the starch branching enzyme 4 (*OsSBE4*) and the sucrose phosphate synthase (*OsSPS*), five primer pairs for the *HvSUT1* gene (HvSUT1m, HvSUT1j, HvSUT1a, HvSUT1b, and HvSUT1c), and three primer pairs for the *nos* terminator (NosTY, NosTYA and NosTh), all present in the T-DNAs, were designed and screened in this study. All selected primer pairs were used in the qPCR assay with 10-fold serial dilutions of the genomic DNA of the A5.1 line (10^0, 10^1, 10^2, 10^3 and 10^4 copies/µl) to determine their relative efficiencies (Table 1). Three primer pairs (SBE4, NostYA and HvSUTj) showed high specificity, generating single PCR products (as seen on the 1.5% agarose gel; Figure S1) and a single peak in the melt curve (Figure 2). Their reproducible amplifications had similar efficiencies close to 100%. The R^2 values for SBE4, HvSUTj and NosTYA primers were 0.997, 0.989 and 0.989, respectively. It was interesting that the three regression lines derived from the qPCR assays with SBE4, HvSUTj and NosTYA primers were parallel. This indicates that the reaction efficiencies of the two primer pairs are very well-matched and comparable within the gDNA template concentration range of 10^0 – 10^4 copies/µl. Two well-matched sets of primer pairs (SBE4/HvSUTj and SBE4/NosTYA) were selected for further analyses.

Based on a logarithmic standard curve generated from the serial dilution of the genomic DNA from the A5.1 transgenic line, an optimized qPCR assay was established to apply in the determination of homozygous and hemizygous plants, as described above. A dilution of 10^3 or 10^2 copies/µl was within the dynamic range of the reaction and considered as the optimal range of starting concentrations of genomic DNA template in the qPCR assay. It was noted that the dilutions with DNA concentration lower than 10^2 copies/µl or higher than 10^3 copies/µl had increased variation between technical replicates. Therefore, a concentration of 10^3 copies/µl of genomic DNA was used as the working concentration in all subsequent experiments.

In 25 offspring of the A5.1 line analysed by this qPCR assay, the prediction was that there were four homozygous plants with the mean $2^{-\Delta\Delta Ct}$ value of 2.00 ± 0.107, 14 hemizygous plants (the mean $2^{-\Delta\Delta Ct}$ value of 1.01 ± 0.041) and seven null plants (the mean $2^{-\Delta\Delta Ct}$ value of 0.00 ± 0.008). These data, when analysed by Chi-square ($\chi^2_{0.05}$) test for goodness of fit, showed no significant difference between the expected segregation ratio (1:2:1) and the observed

TABLE 1 Screening for the specificity, PCR amplification efficiency, slope, and correlation coefficients (R^2) for all primer pairs of the T-DNAs and internal reference genes.

Target gene	Primer pair's name	Specificity (PCR product) number	Melt-curve analysis	Efficiency (%)	Slope	R^2	Reference
Reference genes	Rice Sucrose Phosphate Synthase (*OsSPS*)						
	SPSm	1 band	1 peak	108.5	-3.134	0.992	Mieog et al. (2013)
	SPSj	1 band	No product				
	Rice Starch branching Enzyme (*OsSBE4*)						
	SBE4	**1 band**	**1 peak**	**102.8**	**-3.258**	**0.997**	Wang et al. (2015)
T-DNAs	Barley sucrose transporter (*HvSUT1*)						
	HvSUT1m	1 band	1 peak	120.9	-2.906	0.979	
	HvSUT1j	**1 band**	**1 peak**	**101.9**	**-3.278**	**0.989**	
	HvSUT1a	3 bands	2 peaks	96.4	-3.411	0.987	
	HvSUT1b	3 bands	1 peak	84.1	-3.774	0.988	
	HvSUT1c	1 band	1 peak	122.3	-2.883	0.980	
	Nopaline synthase *(nos)* terminator						
	NosTY	1 band	–	-	-	-	Fletcher (2014)
	NosTYA	**1 band**	**1 peak**	**104.1**	**-3.228**	**0.989**	
	NosTh	1 band	1 peak	88.4	-3.637	0.990	

Specificity of all primer pairs was identified by an endpoint PCR and the products were then visualized on the 1.5% agarose gel (Figure S1). Melt-curve analysis, amplification efficiency, slope and R^2 were identified based on the standard curve conducted by the qPCR assay. Three well-matched primer pairs used in the qPCR assay to determine homozygous plants were highlighted in bold.

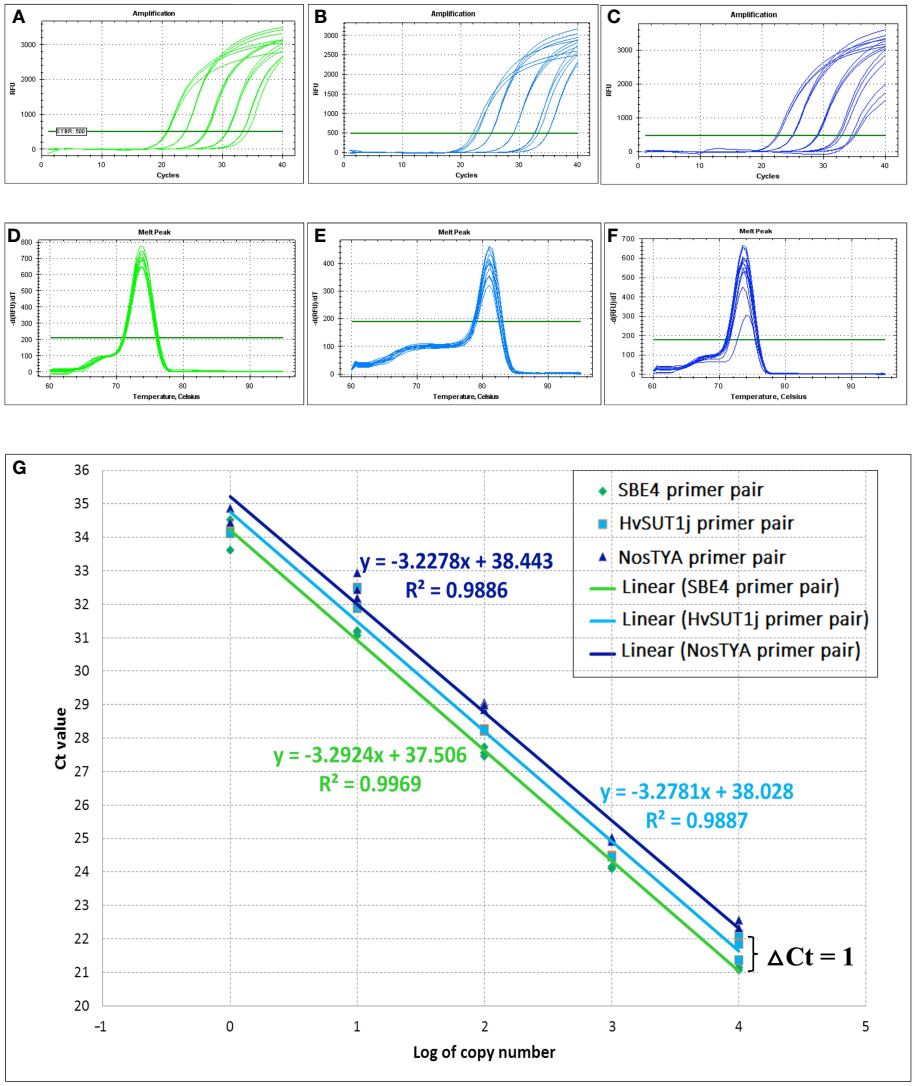


FIGURE 2
Amplification efficiencies of well-matched primer pairs used in the qPCR assay. Amplification curves and melting curves of SBE4 primer pair **(A, D)**, HvSUT1j primer pair **(B, E)** and NosTYA primer pair **(C, F)** for the qPCR assay were conducted in a CFX96^{TM} Real-time PCR detection system (BioRAD). The fluorescent threshold was set at 500 Relative Fluorescent Units (RFU). **(G)** Comparison in amplification efficiency of SBE4, HvSUT1j and NosTYA primers. The genomic DNA from the A5.1 transgenic line with a single T-DNA insert was diluted by 10-fold serial dilution (10^0, 10^1, 10^2, 10^3, and 10^4 copies/µl). The standard curve for each primer pair was made by plotting the average Ct value of three replicates for each dilution against the logarithm of the DNA starting quantity. The green, light blue and dark blue regression lines with their equations and R^2 values are presented for SBE4, HvSUT1j and NosTYA primers.

ratio with a *P* value of 0.583 (Table 2). The results in Table 2 showed that Ct value differences (ΔCt values) for homozygous plants were from 0.008 to 0.17 cycles between Ct values of NosTYA and SBE4, while the ΔCt values of hemizygous plants was from 0.92 to 1.20. For the set of SBE4 and HvSU1j primer pairs, the ΔCt values were more variable with around 0.41 for homozygotes and from 1.20 to 1.46 for hemizygotes (Supplementary Table S4). Moreover, on analysis the standard curves for the two sets of primer pairs, the reactions with NosTYA and SBE4 primer pairs had Ct values for the same dilutions differing by around 1 cycle, while reactions with the HvSUT1j and SBE4 primers had Ct values differing by approximately 0.6 cycles with the same dilutions (Figure 2G). This indicates that the primer set of SBE4 and NosTYA reflected the copy number of the *OsSBE4* reference gene (two copies) doubling the copy number of the T-DNA (one copy) in the diploid genome. In the case of this A5.1 line with a single T-DNA insertion, the quantitative PCR assay reliably distinguishes the zygosity of the transgene utilising the set of primer pairs (NosTYA and SBE4).

To verify the accuracy of zygosity identification by the qPCR assay, T_2 plants of the A5.1 line with a single T-DNA insert were chosen for further segregation analysis. Around 100 progeny from one homozygous plant (No. 6), three hemizygous plants (No. 1, 2 and 5), and one null plant (No. 4), as predicted by the qPCR assay, were analysed for their segregation using hygromycin selection (Supplementary Table S5). Progeny from the plants predicted by

10.3389/fpls.2023.1221790

TABLE 2 Zygosity determination of T_2 plants from the A5.1 transgenic line with a single T-DNA insert.

Plant Number	NosTYA/SBE4					Zygosity	Mean (± SD)
	Ct_{NosT} (± SD)	Ct_{SBE4} (± SD)	ΔCt	ΔΔCt	$2^{-\Delta\Delta Ct}$		
6	**24.14** ± 0.042	**23.98** ± 0.048	0.17	-0.91	1.88	Homozygous	Ct_{Nost}:24.46 ± 0.26.
12	**24.36** ± 0.027	**24.20** ± 0.009	0.16	-0.96	1.95	Homozygous	
13	**24.68** ± 0.034	**24.67** ± 0.062	0.008	-1.09	2.13	Homozygous	$2^{-\Delta\Delta Ct}$: 2.00 ± 0.107
14	**24.67** ± 0.077	**24.59** ± 0.074	0.079	-1.02	2.03	Homozygous	
1	**25.29** ± 0.021	**24.21** ± 0.053	1.08	0.00	1.00	Hemizygous	Ct_{Nost}:25.36 ± 0.26.
2	**25.25** ± 0.012	**24.33** ± 0.061	0.92	-0.16	1.12	Hemizygous	
3	**25.08** ± 0.028	**24.05** ± 0.036	1.03	-0.05	1.03	Hemizygous	$2^{-\Delta\Delta Ct}$: 1.01 ± 0.041
5	**25.07** ± 0.041	**23.99** ± 0.059	1.07	-0.01	1.01	Hemizygous	
7	**25.05** ± 0.068	**24.03** ± 0.015	1.01	-0.07	1.05	Hemizygous	
9	**25.13** ± 0.076	**24.12** ± 0.061	1.01	-0.07	1.05	Hemizygous	
11	**25.80** ± 0.114	**24.70** ± 0.022	1.09	0.00	1.00	Hemizygous	
15	**25.48** ± 0.07	**24.29** ± 0.027	1.19	0.07	0.96	Hemizygous	
16	**25.24** ± 0.029	**24.13** ± 0.009	1.100	0.01	1.00	Hemizygous	
18	**25.29** ± 0.073	**24.16**± 0.057	1.12	0.01	1.00	Hemizygous	
19	**25.60** ± 0.059	**24.51** ± 0.021	1.099	0.00	1.00	Hemizygous	
20	**25.81** ± 0.052	**24.61** ± 0.027	1.20	0.08	0.95	Hemizygous	
23	**25.43** ± 0.048	**24.36** ± 0.026	1.070	-0.03	1.02	Hemizygous	
25	**25.49** ± 0.123	**24.40** ± 0.117	1.086	-0.01	1.01	Hemizygous	
4	**36.14** ± 2.024	**23.93** ± 0.068	12.20	11.12	0.00	Null	Ct_{Nost}:35.69 ± 2.62.
8	**32.39** ± 0.239	**24.13** ± 0.022	8.26	7.18	0.01	Null	
10	**31.59** ± 0.004	**24.92** ± 0.034	6.66	5.57	0.02	Null	$2^{-\Delta\Delta Ct}$: 0.00 ± 0.008
17	**36.85** ± 2.12	**24.73** ± 0.041	12.68	11.60	0.00	Null	
21	**37.91** ± 1.712	**24.58** ± 0.033	13.33	12.23	0.00	Null	
22	**36.96** ± 0.551	**24.50** ± 0.017	12.46	11.36	0.00	Null	
24	**38.02** ± 0.75	**24.31** ± 0.01	13.72	12.59	0.00	Null	
WT	**36.85** ± 0.108	**24.18** ± 0.059	12.68	11.60	0.00	WT	
		Mean $_{Ct\ value}$: 24.32 ± 0.28				4:14:7	P= 0.583^{ns} $\chi^2_{0.05}$ = 1.08

Headings indicate the set of primer pairs used in the qPCR assay. ΔCt, ΔΔCt, and $2^{-\Delta\Delta Ct}$ were calculated based on the formula in the MATERIALS AND METHODS. The $2^{-\Delta\Delta Ct}$ values of homozygous plants should be double that of the hemizygous plants. In the case of the A5.1 transgenic line with single-insert T-DNA, $2^{-\Delta\Delta Ct}$ values of homozygous and hemizygous plants were 2 and 1, respectively. The results were the average and standard deviation (SD) of three replicates from the same plants. All null plants were confirmed to be negative by the endpoint PCR. (ns): no significant difference. WT: non-transgenic rice plant (wild type).

the qPCR to be homozygous were all resistant to hygromycin, whereas progeny from the null plant were all sensitive. Segregation of the T-DNA in the progeny of hemizygous plants showed no significant difference from the expected segregation ratio (3:1 for resistant: sensitive) (Supplementary Table S5). The result of the hygromycin sensitivity assay thus verified the 100% accuracy of the zygosity determined for the A5.1 line as predicted by the qPCR assay.

This optimized qPCR assay using the hemizygous plant from the A5.1 line as a calibrator, was then adapted to determine the zygosity of other transgenic lines. Two transformed lines which contained two T-DNA inserts at the same locus in the rice genome (G1.4 and G3.1) were analysed to determine their zygosity by using the optimized qPCR assay. The qPCR assay identified $2^{-\Delta\Delta Ct}$ values of homozygous plants with four copies as 3.9 to 4.18, and 1.93-1.98 for plants containing two copies (Table 3 and Supplementary Data S2). Only plants with four, two or zero copies of the T-DNA were detected in a 1:2:1 ratio, and the segregation ratios of T-DNA inserts in the G1.4 and G3.1 lines determined by the qPCR assay suggested a single insert locus. The results of the qPCR assay were therefore

TABLE 3 T-DNA segregation analysis of T_1 plants from T_0 plants carrying single insert based on the qPCR assay.

Line	Insert number	Total plants	T_1 segregation (†)	Chi-square (χ^2) value	p value	Percentage of homozygous plant (%)	$2^{-\Delta\Delta Ct}$ value of homozygous plant
DC1.21	1	22	2:12:8	3.45	0.178^{ns}	9.1	3.99-4.13
DC1.1	1	16	4:6:6	1.50	0.472^{ns}	25.0	3.99-4.41
DC1.5	1	24	5:8:11	5.67	0.059^{ns}	20.8	3.91-4.05
DC1.9	1	17	1:10:6	3.47	0.176^{ns}	5.9	3.78
DC1.13	1	38	7:19:12	1.32	0.518^{ns}	18.4	4.04-4.27
SC1.22	1	16	5:8:3	0.50	0.779^{ns}	31.2	1.78-2.17
SC1.21	1	22	3:12:7	1.64	0.441^{ns}	13.6	1.80-2.12
SC1.12	1	8	3:4:1	1.00	0.606^{ns}	37.5	2.01-2.08
SC2.15	1	21	2:14:5	3.19	0.200^{ns}	9.5	1.75-1.87
SC2.31	1	25	2:17:6	4.52	0.104^{ns}	8.0	1.85-2.02
SC2.14	1	8	2:4:2	0.00	1.00^{ns}	25.0	1.93-1.79
SC2.32	1	14	2:7:5	1.46	0.481^{ns}	14.3	1.79-1.81
G1.4	2	11	2:4:5	2.45	0.293^{ns}	18.1	3.96-4.01
G3.1	2	5	2:1:2	1.80	0.407^{ns}	40.0	3.9-4.18
SC1.16	2	10	3: 0: 5: 0: 2	174.67	0.00*	30.0	3.79-4.18
SC1.13	2	11	1: 0: 5: 0: 5	298.84	0.00*	9.1	3.95
	Total	268			Average (%)	19.72	

The copy number of the T-DNA was estimated by the qPCR assay using the set of NostYA and SBE4 primers. χ^2 test ($\chi^2_{0.05} = 5.99$; dF=2) for goodness of fit was applied to compare whether there was a significant difference between the observed segregation and expected segregation at $p < 0.05$. (†): In the case of a single insert, the segregation of T-DNA is 1: 2: 1 for 0: 1: 2 copies, respectively. The independent segregation of two T-DNA inserts in case of two-insert events is expected to be 1:4:6:4:1 for 4:3:2:1:0 copies, respectively. The hemizygous, single-insert plant from the A5.1 line was used as a calibrator. (*): significant difference; (ns): no significant difference.

consistent with the Southern blot analyses. The use of the qPCR assay to confidently detect homozygous plants where two T-DNAs had integrated at the one locus is a powerful and useful application of this tool.

Application of the qPCR method

Having optimized the qPCR assay, this method was successfully applied to fast-track the identification of T-DNA copy number and zygosity of transgenic Indica rice plants in the T_1 generation. In this experiment, three populations of transgenic Indica rice lines carrying T-DNAs with a single transgene of the *HvSUT1*, the rice nicotinanamine synthase (*OsNAS2*), and both the transgenes (*HvSUT1* and *OsNAS2*) were developed and designated as SC1, SC2 and DC1 lines, respectively (Figures 1C–E). All the transgenic lines contained the same *nos* sequence element as the A5.1 line (Figure 1B), so the qPCR assay with the same primer pairs for T-DNAs and reference gene could be used without any adaptation. The hemizygous plant from the A5.1 line was used as a calibrator for all qPCR runs. To achieve the fastest route for developing a homozygous line, we could detect single-insert lines in the T_0 generation.

The T-DNAs in all the T_0 putative transgenic lines exist in a hemizygous state. For single-transgene T-DNA lines (SC1 and SC2), Ct values (~ 25) of the NosTYA primer pair (Figures 3D, E) were higher by approximately 1, compared to the Ct values (~ 24) of the reference SBE4 primer pair (Figures 3A, B). The results were comparable to the hemizygous calibrator. Based on the $2^{-\Delta\Delta Ct}$ value of each putative transgenic line in the T_0 generation, the copy number of T-DNA inserts was predicted, with one indicating a hemizygous single insert, two for two hemizygous inserts, etc. The results are summarized in Figures 3G, H, with the putative transgenic events showing a single insertion of the T-DNAs as the most frequent, with seven of the 21 SC1 lines (33.33%) and ten of the 21 SC2 lines (47.62%). The transgenic lines likely to contain a single insertion per genome in the SC1 and SC2 lines had a $2^{-\Delta\Delta Ct}$ value of 0.8 - 1.0, and 0.7 - 1.1, respectively. For T_0 transgenic lines carrying the DC1 T-DNA, Ct values of the NosTYA primer pair were similar to that of the reference SBE4 primer pair, around 24, because in this case there were two *nos* terminator regions in the T-DNA (Figures 3C, F). The $2^{-\Delta\Delta Ct}$ value was ~ 2 for single-insert lines, ~ 4 for two-insert lines, etc. The result, shown in Figure 3I, indicates eleven of the 30 DC1 transgenic lines (36.67%) were determined to carry a single insertion of the DC1 T-DNA with $2^{-\Delta\Delta Ct}$ values of 1.8 – 2.2.

The T_0 lines carrying a single insert for the SC1, SC2 and DC1 T-DNAs were grown and used to screen for homozygous plants in the T_1 generation. The theoretical segregation of the T-DNA in single-insert lines would be consistent with a Mendelian ratio of 1:2:1 for

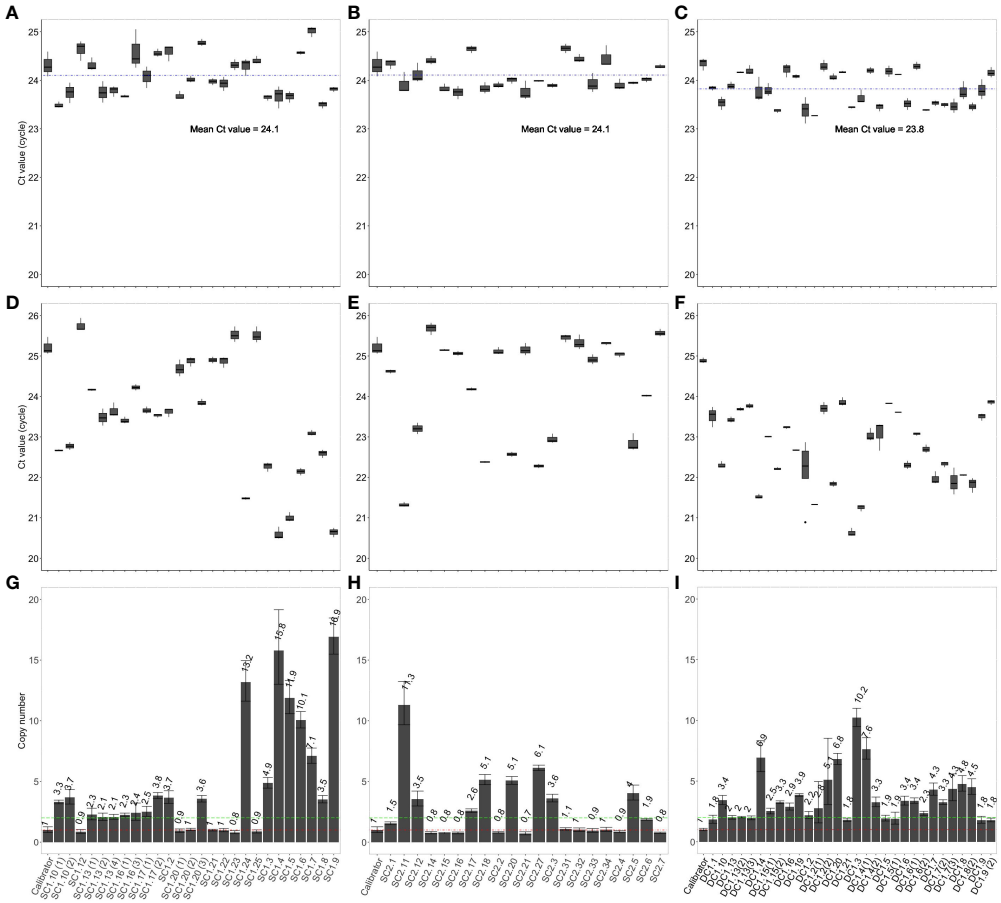


FIGURE 3
Copy number determination in T_0 transgenic rice plants using the qPCR assay. Average Ct values of SBE4 **(A)** and NosTYA **(D)** and copy number of T-DNAs **(G)** in T_0 transgenic rice plants carrying the SC1 T-DNA. Average Ct values of SBE4 **(B)** and NosTYA **(E)** and copy number of T-DNAs **(H)** in T_0 transgenic rice plants carrying the SC2 T-DNA. Average Ct values of SBE4 **(C)** and NosTYA **(F)** and copy number of T-DNAs **(I)** in T_0 transgenic rice plants carrying the DC1 T-DNA with two stacked genes. The qPCR assay was adapted without any modification. The A5.1 transgenic line was used as a calibrator. The blue lines represent the mean Ct values of the SBE4 primer pair in all SC1 and SC2 lines (24.1) and DC1 lines (23.8). Red and green lines represent 1 and 2 copies of T-DNAs, respectively.

homozygous, hemizygous and null plants. Seeds from 13 independent transgenic events in the T_0 generation, predicted to have a single insertion by qPCR, were grown to determine zygosity in the T_1 generation using qPCR (Table 3). A total of 37 homozygous T_1 plants were predicted from the qPCR assay; 19 homozygous plants from five independent transformation events carrying the DC1 T-DNA with the $2^{-\Delta\Delta Ct}$ values of 3.78 - 4.41, 11 homozygous plants from three independent SC1 transgenic lines with the $2^{-\Delta\Delta Ct}$ values of 1.78 - 2.17 and seven homozygous plants from four independent SC2 transgenic lines with the $2^{-\Delta\Delta Ct}$ values of 1.75 – 2.02. A Chi-squared ($\chi^2_{0.05}$) analysis was used to analyse the segregation ratios of the T_1 progenies based on the qPCR results. In all cases, there was no significant difference between the observed ratios and that expected for Mendelian segregation of a single gene. The frequencies of homozygous plants carrying the SC1 T-DNA (23.9%) were much closer to the expected frequency of 25%. There were lower frequencies in the SC2 lines (only 10.3%) and in the DC1 lines (16.2%). There was no T_1 progeny plant from the single-insert lines with more than two copies identified by the qPCR assay. For the SC1 two-insert lines, homozygous plants were identified with $2^{-\Delta\Delta Ct}$ values from 4.07 - 4.14, compared to 1 for the hemizygous calibrator.

To verify the reliability and accuracy of the qPCR assay to determine zygosity and number of T-DNA insertions, 105 T_2 progeny from eleven homozygous single-insert plants and one homozygous double-insert line (shown in Table 3) were analyzed for the presence of the T-DNA by end-point PCR using the specific primer pairs for the transgenes (Supplementary Table S2). The results (Table 4) indicate that all progeny from the plants predicted to be homozygous for a single T-DNA insert were positive for the presence of the T-DNA by end-point PCR, demonstrating 100% accuracy for identifying homozygous plants attained from this qPCR assay in the T_1 generation. A similar result was also achieved for the homozygous, two-insert transgenic plants determined by the qPCR. Therefore, all materials from those homozygous plants and their progenies could be confidently used for further phenotypic, molecular, and physiological analysis.

TABLE 4 Summarizing end-point PCR analysis of T_2 plants from the homozygous T_1 plants.

Line	Insert number	Number of plants detected	Number of PCR positive plants	Percentage (%)
DC1.1	1	10	10	100
DC1.5	1	8	8	100
DC1.9	1	9	9	100
DC1.13	1	10	10	100
SC1.22	1	10	10	100
SC1.21	1	6	6	100
SC1.12	1	10	10	100
SC1.13	2	9	9	100
SC2.32	1	7	7	100
SC2.15	1	10	10	100
SC2.14	1	7	7	100
SC2.31	1	9	9	100
Total	–	105	105	100

All T_2 progeny from homozygous T_1 plants were positively confirmed for the presence of T-DNAs by endpoint PCR.

Discussion

Ideally, identifying single-insert T_0 transgenic lines can greatly assist in the development of homozygous T_1 lines. However, determining homozygotes is challenging because neither end-point PCR nor Southern blotting can clearly and reliably differentiate between homozygous and hemizygous plants at this point in the experimental timeline. Therefore, genetic analysis of T_2 progeny is required, which is time-consuming and space and labour-intensive. Alternatively, the qPCR assay can determine copy number and zygosity (Ingham et al., 2001; German et al., 2003; Bubner and Baldwin, 2004; Bubner et al., 2004; Mieog et al., 2013; Wang et al., 2015), however these studies have limitations in reliability and universality. A standardized and universal qPCR protocol to identify homozygous plants accurately, reliably, and rapidly is needed and is presented here.

By using the comparative $2^{-\Delta\Delta Ct}$ method, the qPCR assay was optimized to provide an accurate, reliable, and powerful tool to identify single insert T_0 plants and thus fast track the development of homozygous transgenic lines. The comparative $2^{-\Delta\Delta Ct}$ method requires an endogenous reference gene and equal PCR amplification efficiencies for the reference gene and the T-DNA construct (Bubner and Baldwin, 2004). In this study, the *OsSBE4* gene was used as an endogenous reference, and the *nos* terminator for the T-DNA, were targeted to design a specific, well-matched pair of primers to achieve equal amplification efficiencies. The genomic DNA from the A5.1 transgenic line with a single T-DNA insert, as confirmed by Southern blot analysis, contains one copy of the *nos* terminator and two copies of the *OsSBE4* reference gene (Mizuno et al., 2001). The A5.1 line was used as starting DNA to analyse the PCR efficiencies of the SBE4 and NosTYA primer pairs by plotting the average Ct values for each dilution and fitting a standard curve (Figure 2G). These standard curves showed close to 100% amplification efficiency in each case. The difference in Ct values at each dilution between the SBE4 and NosTYA primers was consistently one cycle and parallel. This reflects the difference in copy number of the T-DNA (one copy) and the endogenous reference gene (two copies). From these results, the optimized protocol of the qPCR assay used in this study was established with standardized procedures for dynamic DNA concentration, well-matched primer pairs, real-time PCR conditions, and the $2^{-\Delta\Delta Ct}$ method. The qPCR assay was utilized to screen the progeny of hemizygous, single T-DNA insert lines. The results indicated 100% accuracy in determining homozygous (two copies), hemizygous (one copy), and null (zero copy) plants. Two major factors contributed to the high accuracy of the qPCR assay achieved in this study. The first is an internal reference gene with a single copy per haploid genome. Many previous reports did not determine the copy number of the reference gene (Bubner and Baldwin, 2004), and this could affect the T-DNA copy number estimate. In practice, a reference gene when present in multiple copies per haploid genome would be more complicated to amplify and estimate by real-time PCR using SYBR green. It is therefore of no surprise that many previous reports were less successful in accurately measuring T-DNA insert copy number (Bubner and Baldwin, 2004). Meanwhile, Mieog et al. (2013) and Wang et al. (2015) reported success in determining T-DNA copy number when the reference genes were present in a single copy per haploid genome. However, their Ct values of the reference genes and T-DNAs were variable in all plants with identical zygosity. In this study, the qPCR assay showed very little variation in the Ct values of the reference gene and T-DNAs. The robust and consistent results of the Ct values in every plant or sample was a direct consequence of our well-optimized and standardized protocol.

The second factor contributing to the accuracy and reliability of the assay was the Ct values of the endogenous gene (two copies) which in all the plants was consistently around 24 with a low standard deviation (0.28), and very low standard deviation (<0.14) between three replicates, as is required for copy number and

zygosity determinations (Bubner et al., 2004). We conclude that the low variation in Ct values of the endogenous gene was easily achieved because of the high quality and consistent concentration of starting DNA template (10^3 copies/µl or 0.426 ng/µl). In the previous reports in which Ct values of reference genes in samples varied significantly from 19 to 21 (Bubner and Baldwin, 2004; Wang et al., 2015), there were no standardized qPCR protocols. Furthermore, in our study, the Ct values of the *nos* terminator in the homozygous plants were all around 24, similar to the Ct values of the endogenous gene, but higher by 1 cycle in the hemizygous plants (around 25). This is more likely to reflect the copy number of the T-DNAs in the single-insert plants. Based on the Ct values of the reference gene and the T-DNA, the zygosity of single insert transgenic plants could then be accurately predicted. This level of consistency was not seen in data of the previous reports (Bubner and Baldwin, 2004; Wang et al., 2015).

Along with the high accuracy and reliability, the assay was demonstrated to be universal and repeatable for copy number and zygosity determination in transgenic rice. The assay was successfully adapted for determining multiple T-DNAs with the same *nos* terminator in two rice varieties (Indica and Japonica) at the same time and without any modifications. The NosTYA primer pair was used to detect the *nos* terminator which is a commonly used regulatory element in genetically modified of crops (Wu et al., 2014) and the SBE4 primer pair was used to amplify the endogenous reference gene, which is present as two copies per rice genome and well conserved in different rice including the most popular Indica and Japonica varieties (Mizuno et al., 2001; Wang et al., 2015). The qPCR-based methods established in the previous reports required several modifications for wider application, namely primer design for T-DNAs or reference genes, the selection of the appropriate endogenous gene, and the optimisation of the qPCR assay conditions. These modifications are crucial to achieve the level of accuracy reported here.

We demonstrated here, a fast-tracking strategy for high throughput development of single-insert homozygous T_1 lines derived from single-insert T_0 lines, as only two generations are required to identify at least one homozygous plant, from which 100% of the T_2 progeny were homozygous. In addition, the qPCR assay has the potential to reduce cost in large-scale screening of homozygous plants. First, this qPCR method required a small amount of plant tissues (~100 mg) of T_1 progeny to obtain enough genomic DNA (1.065 ng) for each qPCR reaction, thus we determined homozygous T_1 transgenic Indica rice plant within 4 weeks after germination, thus reducing space and resources required for growing numerous hemizygous and null T_1 progeny (75% in theory, compared to approximately 80.28% in this study). Second, the proposed protocol using SYBR Green qPCR, provides a simple and transferable assay to molecular breeding and transgenic research being much cheaper and easier to use than the TaqMan assay. Additionally, the data were unchanged when the qPCR reactions were set up in a 10 µl final volume, instead of 20 µl reactions, further reducing the experimental cost by a half. Finally, cheaper end-point PCR was used for initial screening of transgenic plants from the T_1 population to remove all null plants. This reduced the experimental cost of the SYBR Green qPCR ReadyMix by a further 25%.

For T_0 lines with two inserts located at different loci, multiple generations may be required to achieve single-copy homozygous plants (Mieog et al., 2013). The qPCR assay could identify two copies of the T-DNA but would be more difficult to distinguish between homozygous single insert progeny and hemizygous two-insert progeny. Although transgenic lines with a single insertion are preferred, two inserts of the T-DNA can still be useful in genetic studies. The qPCR assay was able to detect homozygous, two-insert plants containing four copies. In the case of the G1.4 and G3.1 lines containing two T-DNA inserts in a tandem inverted orientation at one locus (Figures 1F, G), a good correlation was found between the results of the qPCR and Southern blot assays. This result indicates that the qPCR assay was accurate and reliable and thus we successfully applied the assay to develop homozygous plants carrying the two-gene cassette (the DC1 T-DNAs) using a single reaction. In contrast, Wang et al. (2015) needed three separate reactions with three different primer pairs to detect homozygous lines stacked with three different gene cassettes at different loci in rice. The qPCR protocol used here with universal primer pairs (NosTYA) was employed to detect the homozygous lines stacked with two different genes using the same terminator sequence. This demonstrates the wide application of this qPCR assay to transgenic plant research.

Our method obtained accurate measurements of T-DNA copy number in rice that are very close to integer values. Such accuracy was very similar or even better that that achieved by the ddPCR method using a more sophisticated droplet digital PCR systems (Głowacka et al., 2016; Xu et al., 2016; Collier et al., 2017; Giraldo et al., 2019). For instance, our study reported that T-DNA copy number using the proposed qPCR method was 0.94 - 1.11 and 1.85 – 2.07 for single T-DNA copy and two T-DNA copy events, respectively. In our larger experiment, the measurements of T-DNA copy were 0.7 – 1.1 and 1.75 – 2.17, respectively. Meanwhile, Collier et al. (2017) reported 0.83 – 1.57 for single T-DNA copy, and 1.63 – 2.97 for two T-DNA copies, and Głowacka et al. (2016) showed 0.94 - 1.12 and 1.80 - 2.26 for single T-DNA copy and two T-DNA copy events, respectively. Recently only two reports, of Głowacka et al. (2016) and Xu et al. (2016), have demonstrated the successful used of the ddPCR to identify of homozygous plants. Unfortunately, in these cases no additional analysis was carried out to verify the segregation of the T-DNA in the next generation and thus the accuracy of the method could not be verified. In comparison to qPCR, the ddPCR method is more complicated than that proposed here. For example, the use of restriction digested genomic DNA containing two T-DNA copies in a tandem inverted orientation at one locus, which is relatively common in transgenic plants (Tzfira et al., 2004; De Buck et al., 2009), presents a challenge for the ddPCR method that is not the case in our method. In the study of Collier et al. (2017), one R5-24 transgenic rice line was estimated to contain a single T-DNA insert using the ddPCR method, but two T-DNA inserts using Southern blot. In the case of our study, the qPCR and Southern blot analysis of the G1.4 and G3.1 transgenic lines were consistent with a tandem inverted T-

DNA insertion event. Furthermore, the ddPCR method has a much higher experimental cost and is more time-consuming than our qPCR method because the ddPCR relies on an initial restriction digestion of genomic DNA, expensive labelled probes specific to transgenes, and a costly droplet digital PCR system for T-DNA copy measurement (Cai et al., 2021). In addition, the primer pairs and labelled probes used in the ddPCR assay are specific to the T-DNA and thus would be time-consuming and expensive to optimize. Further, if multiple T-DNAs were investigated using the ddPCR method, the cost and time factors would be amplified.

To conclude, we present here an improved and efficient qPCR assay to identify homozygous transgenic plants confidently and economically. This assay is suitable for lower budget laboratories that are involved in transgenic research and could also be applied to a variety of transgenic plant species carrying T-DNAs with the same regulatory elements.

Data availability statement

Sequence data may be found in the NCBI's Genbank under accession numbers: the pIPKb001 vector (EU161567), the pIPKb003 vector (EU161569), *HvSUT1* (AJ272309), rice *Glb-1* promoter (AY427575), rice constitutive *Act-1* promoter (S44221.1), *OsSPS* (AP003437), *OsSBE4* (GQ150932), *Nos* terminator (MK078637.1), and rice *GluA2* (EU264103).

Author contributions

HT and CS optimized the protocol. HT designed and performed the experiments. HT and MH contributed to plant transformation. HT completed statistical analysis of the data. PA, YS, CJ, and JCRS assisted HTT in conceiving the project. HT wrote the manuscript in consultation with PA. HT, PA, CS, YS, CJ, and JS discussed the results and contributed to the final manuscript.

Acknowledgments

Research was supported by Flinders University Research Scholarship and Australia Award Scholarship. The authors gratefully acknowledge Dr. Julien Pierre Bonneu and Dr. Alex Johnson (The University of Melbourne) and Dr. Hans Weber and Dr. Jochen Kumlehn (IPK, Germany) for kindly providing the cDNA of *OsNAS2* and *HvSUT1* and the pIPKb001 and pIPKb003 vectors. I would like to acknowledge Dr Larissa Chirkova at the ACPFG (Waite campus, Adelaide University) for her assistance with hybridization and detection in Southern blot analyses.

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.2023.1221790/full#supplementary-material

References

Applied Biosystems (2003). *Creating standard curves with genomic DNA or plasmid DNA templates for use in quantitative PCR* (F Hoffmann-La Roche Ltd.: Applied Biosystems Carlsbad) California. Available at: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_042486.pdf.

Baric, S., Kerschbamer, C., and Via, J. D. (2006). TaqMan real-time PCR versus four conventional PCR assays for detection of apple proliferation phytoplasma. *Plant Mol. Biol. Rep.* 24 (2), 169–184. doi: 10.1007/BF02914056

Bio-Rad Laboratories (2006). *Real-time PCR applications guide* (Bio-Rad Laboratories Hercules, CA: Life Science Group). Available at: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_5279.pdf.

Bubner, B., and Baldwin, I. T. (2004). Use of real-time PCR for determining copy number and zygosity in transgenic plants. *Plant Cell Rep.* 23 (5), 263–271. doi: 10.1007/s00299-004-0859-y

Bubner, B., Gase, K., and Baldwin, I. (2004). Two-fold differences are the detection limit for determining transgene copy numbers in plants by real-time PCR. *BMC Biotechnol.* 4, 14. doi: 10.1186/1472-6750-4-14

Cai, Y.-M., Dudley, Q. M., and Patron, N. J. (2021). Measurement of transgene copy number in plants using droplet digital PCR. *Bio-protocol* 11 (13), e4075–e4075. doi: 10.21769/BioProtoc.4075

Cai, Y.-M., Kallam, K., Tidd, H., Gendarini, G., Salzman, A., and Patron, N. J. (2020). Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res.* 48 (21), 11845–11856. doi: 10.1093/nar/gkaa682

Collier, R., Dasgupta, K., Xing, Y. P., Hernandez, B. T., Shao, M., Rohozinski, D., et al. (2017). Accurate measurement of transgene copy number in crop plants using droplet digital PCR. *Plant J.* 90 (5), 1014–1025. doi: 10.1111/tpj.13517

De Buck, S., Podevin, N., Nolf, J., Jacobs, A., and Depicker, A. (2009). The T-DNA integration pattern in Arabidopsis transformants is highly determined by the transformed target cell. *Plant J.* 60 (1), 134–145. doi: 10.1111/j.1365-313X.2009.03942.x

Fletcher, S. J. (2014). qPCR for quantification of transgene expression and determination of transgene copy number. *Methods Mol. Biol.* 1145, 213–237. doi: 10.1007/978-1-4939-0446-4_17

Gadaleta, A., Giancaspro, A., Cardone, M. F., and Blanco, A. (2011). Real-time PCR for the detection of precise transgene copy number in durum wheat. *Cell. Mol. Biol. Lett.* 16 (4), 652–668. doi: 10.2478/s11658-011-0029-5

German, M. A., Kandel-Kfir, M., Swarzberg, D., Matsevitz, T., and Granot, D. (2003). A rapid method for the analysis of zygosity in transgenic plants. *Plant Sci.* 164 (2), 183–187. doi: 10.1016/S0168-9452(02)00381-3

Giancaspro, A., Gadaleta, A., and Blanco, A. (2017). Real-time PCR for the detection of precise transgene copy number in wheat. *Methods Mol. Biol.* 1679, 251–257. doi: 10.1007/978-1-4939-7337-8_15

Giraldo, P. A., Cogan, N. O., Spangenberg, G. C., Smith, K. F., and Shinozuka, H. (2019). Development and application of droplet digital PCR tools for the detection of transgenes in pastures and pasture-based products. *Front. Plant Sci.* 9, 1923. doi: 10.3389/fpls.2018.01923

Głowacka, K., Kromdijk, J., Leonelli, L., Niyogi, K. K., Clemente, T. E., and Long, S. P. (2016). An evaluation of new and established methods to determine T-DNA copy number and homozygosity in transgenic plants. *Plant Cell Environ.* 39 (4), 908–917. doi: 10.1111/pce.12693

Ingham, D. J., Beer, S., Money, S., and Hansen, G. (2001). Quantitative real-time PCR assay for determining transgene copy number in transformed plants. *Biotechniques* 31 (1), 132–140. doi: 10.2144/01311rr04

Jiang, L., Yang, L., Zhang, H., Guo, J., Mazzara, M., Van den Eede, G., et al. (2009). International collaborative study of the endogenous reference gene, sucrose phosphate synthase (*SPS*), used for qualitative and quantitative analysis of genetically modified rice. *J. Agric. Food Chem.* 57 (9), 3525–3532. doi: 10.1021/jf803166p

Li, Z., Hansen, J., Liu, Y., Zemetra, R., and Berger, P. (2004). Using real-time PCR to determine transgene copy number in wheat. *Plant Mol. Biol. Rep.* 22 (2), 179–188. doi: 10.1007/BF02772725

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

Mason, G., Provero, P., Vaira, A. M., and Accotto, G. P. (2002). Estimating the number of integrations in transformed plants by quantitative real-time PCR. *BMC Biotechnol.* 2, 20. doi: 10.1186/1472-6750-2-20

Mieog, J., Howitt, C., and Ral, J.-P. (2013). Fast-tracking development of homozygous transgenic cereal lines using a simple and highly flexible real-time PCR assay. *BMC Plant Biol.* 13, 71. doi: 10.1186/1471-2229-13-71

Mizuno, K., Kobayashi, E., Tachibana, M., Kawasaki, T., Fujimura, T., Funane, K., et al. (2001). Characterization of an isoform of rice starch branching enzyme, RBE4, in developing seeds. *Plant Cell Physiol.* 42 (4), 349–357. doi: 10.1093/pcp/pce042

Schmidt, M., and Parrott, W. (2001). Quantitative detection of transgenes in soybean [*Glycine max* (L.) Merrill] and peanut (*Arachis hypogaea* L.) by real-time polymerase chain reaction. *Plant Cell Rep.* 20 (5), 422–428. doi: 10.1007/s002990100326

Sequencing Project International Rice. (2005). The map-based sequence of the rice genome. *Nature* 436 (7052), 793–800. doi: 10.1038/nature03895

Shou, H., Frame, B., Whitham, S., and Wang, K. (2004). Assessment of transgenic maize events produced by particle bombardment or *Agrobacterium*-mediated transformation. *Mol. Breed.* 13 (2), 201–208. doi: 10.1023/B:MOLB.0000018767.64586.53

Song, P., Cai, C., Skokut, M., Kosegi, B., and Petolino, J. (2002). Quantitative real-time PCR as a screening tool for estimating transgene copy number in WHISKERS™-derived transgenic maize. *Plant Cell Rep.* 20 (10), 948–954. doi: 10.1007/s00299-001-0432-x

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98 (3), 503–517. doi: 10.1016/s0022-2836(75)80083-0

Tzfira, T., Li, J., Lacroix, B. t., and Citovsky, V. (2004). *Agrobacterium* T-DNA integration: Molecules and models. *Trends Genet.* 20 (8), 375–383. doi: 10.1016/j.tig.2004.06.004

Wang, X., Jiang, D., and Yang, D. (2015). Fast-tracking determination of homozygous transgenic lines and transgene stacking using a reliable quantitative real-time PCR assay. *Appl. Biochem. Biotechnol.* 175 (2), 996–1006. doi: 10.1007/s12010-014-1322-3

Weiss, A. S. (1992). Southern transfer and hybridization: A class experiment. *Biochem. Educ.* 20 (4), 231–233. doi: 10.1016/0307-4412(92)90202-W

Wu, Y., Wang, Y., Li, J., Li, W., Zhang, L., Li, Y., et al. (2014). Development of a general method for detection and quantification of the P35S promoter based on assessment of existing methods. *Sci. Rep.* 4 (1), 7358. doi: 10.1038/srep07358

Xu, X., Peng, C., Wang, X., Chen, X., Wang, Q., and Xu, J. (2016). Comparison of droplet digital PCR with quantitative real-time PCR for determination of zygosity in transgenic maize. *Transgenic Res.* 25 (6), 855–864. doi: 10.1007/s11248-016-9982-0

Yang, L., Ding, J., Zhang, C., Jia, J., Weng, H., Liu, W., et al. (2005). Estimating the copy number of transgenes in transformed rice by real-time quantitative PCR. *Plant Cell Rep.* 23 (10-11), 759–763. doi: 10.1007/s00299-004-0881-0

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 December 2023
DOI 10.3389/fpls.2023.1296197

OPEN ACCESS

EDITED BY
Yuri Shavrukov,
Flinders University, Australia

REVIEWED BY
Mehraj Abbasov,
Azerbaijan National Academy of Sciences, Azerbaijan
Manish Kumar Vishwakarma,
Borlaug Institute for South Asia (BISA), India

*CORRESPONDENCE
Yonghong Gao
gaoyh@xaas.ac.cn

RECEIVED 18 September 2023
ACCEPTED 14 November 2023
PUBLISHED 01 December 2023

CITATION
Ding Y, Fang H, Gao Y, Fan G, Shi X, Yu S, Ding S, Huang T, Wang W and Song J (2023) Genome-wide association analysis of time to heading and maturity in bread wheat using 55K microarrays. *Front. Plant Sci.* 14:1296197. doi: 10.3389/fpls.2023.1296197

Genome-wide association analysis of time to heading and maturity in bread wheat using 55K microarrays

Yindeng Ding[1], Hui Fang[1], Yonghong Gao[1]*, Guiqiang Fan[1], Xiaolei Shi[2], Shan Yu[3], Sunlei Ding[2], Tianrong Huang[1], Wei Wang[4] and Jikun Song[5]

[1]Institute of Grain Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, Xinjiang, China, [2]Institute of Crop Variety Resources, Xinjiang Academy of Agricultural Sciences, Urumqi, Xinjiang, China, [3]College of Agriculture, Xinjiang Agricultural University, Urumqi, Xinjiang, China, [4]Department of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang, China, [5]Cotton Research Institute, Chinese Academy of Agricultural Sciences, Anyang, China

To investigate the genetic mechanisms underlying the reproductive traits (time to flowering and maturity) in wheat and identify candidate genes associated, a phenotypic analysis was conducted on 239 wheat accessions (lines) from around the world. A genome-wide association study (GWAS) of wheat heading and maturity phases was performed using the MLM (Q+K) model in the TASSLE software, combined with the Wheat 55K SNP array. The results revealed significant phenotypic variation in heading and maturity among the wheat accessions across different years, with coefficients of variation ranging from 0.96% to 1.97%. The phenotypic data from different years exhibited excellent correlation, with a genome-wide linkage disequilibrium (LD) attenuation distance of 3 Mb. Population structure analysis, evolutionary tree analysis, and principal component analysis indicated that the 239 wheat accessions formed a relatively homogeneous natural population, which could be divided into three subgroups. The GWAS results identified a total of 293 SNP marker loci that were significantly associated with wheat heading and maturity stages ($P \leq 0.001$) in different environments. Among them, nine stable SNP marker loci were consistently detected in multiple environments. These marker loci were distributed on wheat chromosomes 1A、1B、2D、3A、5B、6D and 7A. Each individual locus explained 4.03%-16.06% of the phenotypic variation. Furthermore, through careful analysis of the associated loci with large phenotypic effect values and stable inheritance, a total of nine candidate genes related to wheat heading and maturity stages were identified. These findings have implications for molecular marker-assisted selection breeding programs targeting specific wheat traits at the heading and maturity stages. In summary, this study conducted a

comprehensive GWAS of wheat heading and maturity phases, revealing significant associations between genetic markers and key developmental stages in wheat. The identification of candidate genes and marker loci provides valuable information for further studies on wheat breeding and genetic improvement targeted at enhancing heading and maturity traits.

KEYWORDS

wheat, heading, maturity, GWAS, candidate genes

Introduction

Wheat is one of the most important grain crops in the world and the third largest grain crop in China, and its yield plays a crucial role in China's food security (Cheng et al., 2023). Different reproductive stages can reflect the growth and development rate and stability of wheat, and have a close relationship with maturity, yield and disease resistance of wheat, and have a great influence on the planting environment, regional planning, and the selection of varieties and cultivation and management measures of wheat, which has been a long-term concern of breeders (Hoogendoorn, 1985; Xu et al., 2022). Quality seedling emergence plays an important role in the yield of the subsequent crop, and the appropriate heading stage can ensure high and stable yield of wheat. The maturity stage of wheat controls the growth cycle of the crop's reproductive period and provides a new direction for the selection of accessions. In addition, unfavorable climate change, shortage of natural resources and the threat of pathogenicity of pests and diseases in recent years have provided new challenges to wheat yield and quality as well as to the growth cycle of wheat (Morales et al., 2020).

A large number of studies have shown that reproduction period is controlled by multiple genes and a typical quantitative trait and that is susceptible to environmental influences. Mining gene loci associated with reproductive traits is an important basis for molecular marker-assisted breeding and interpretation of gene effects (Mackay and Powell, 2007). In recent years, with the emergence of new sequencing technologies, the reduction of sequencing costs and the release of wheat reference genome information, the development and application of wheat SNP microarrays have become more popular, and a large number of genetic loci controlling reproductive traits (time to flowering and maturity) have been excavated at home and abroad.

Nine QTLs located on 2D, 3B and 3D chromosomes were detected, with the highest contribution rate of 22.91% to heading stage (Song et al., 2006). A total of 9 heading stage related QTLs were detected on chromosomes 1B, 2B, 4B, 5A, 5B and 7B, which could explain 4. 55%-13. 40% of phenotypic variation (Wang et al., 2020). Wheat heading stage conformed to the genetic model of a pair of main genes + multiple genes through a multi-generation joint analysis (Wang et al., 2007). In recent years, there have been more reports on the studies of analyzing and locating QTL for wheat heading stage. A series of heading QTLs located on chromosomes 1A, 2B, 4D, 4BS, 5AL, 5DL, 7BS, 2D, 5A, 4A, 2A, and 2D were detected by DH populations (Sourdille et al., 2000). Flowering is a more water-sensitive period in wheat, which directly affects wheat yield. Two flowering QTLs located on chromosomes 1B and 1D under drought stress were detected through DH population, explaining 12.35% and 10.79% of the phenotypic variation, respectively, and two flowering QTLs located on chromosomes 1D and 5B under normal irrigation conditions, explaining 9.11% and 9.65% of the phenotypic variation, respectively (Yan et al., 2015). Five stable QTLs were detected on chromosomes 2A, 5B, 6B, 7A and 7D by genotyping RIL population with 90K chip, among which *QHd.cau-7D* can explain 29.35%-41.96% of phenotypic variation (Chen et al., 2020). Although a large number of wheat heading and anthesis loci have been reported, few of them have been used for breeding selection, and the loci controlling heading and anthesis differ from each other and the mechanism of inheritance is complex, with materials of different genetic backgrounds carrying different resistance factors. In this study, 239 wheat accessions (lines) were identified at heading and maturity stages, and genome-wide linkage analyses were carried out with SNP markers, in order to provide references for the study of genetic mechanisms of heading and maturity stages and molecular breeding of wheat.

Materials and methods

Materials

A total of 239 natural populations consisting of wheat breeding accessions and foreign introduced accessions promoted in winter wheat areas of China were used as test materials (Table S1). Among them, 213 materials were from China, including 6 from Anhui, 34 from Beijing, 19 from Hebei, 19 from Henan, 5 from Jiangsu, 11 from Shandong, 1 from Shanxi, 3 from Shaanxi, 1 from Tianjin and 115 from Xinjiang. There were 26 foreign materials, including 25 from the USA and 1 from Ukraine. All these materials could grow and develop normally in the test site.

Experimental design

The experiment was conducted from September 2019 to July 2022 at Zepu Breeding Base (77°16'17.22 "N, 38°11'21.65 "E) of Xinjiang Academy of Agricultural Sciences (XAS) and in June 2021 at Anningqu Base (43°58'53.38 "N, 87°30'17.72 "E) of XAS. Where Xinjiang Zepu in 2020 was E1, Xinjiang Zepu in 2021 was E2, Xinjiang Anningqu in 2021 was E3, and Xinjiang Zepu in 2022 was

E4. The three-year experiments at the four environments were conducted in a randomized block design with two rows of each material, row length of 1 m, row spacing of 0.2 m, and seeding of about 525 grains per square meter in a north-south row orientation. Light management, water and fertilizer management, water management and other field management were carried out according to normal management, each with three replications. The farming conditions and production conditions of each replication were the same.

Determination of main reproductive traits (time to flowering and maturity)

Wheat plant samples were collected during several developmental stages, at seedling emergence, heading stage and maturity stage respectively. When more than half of the first true leaves of a variety were exposed to 2-3 cm above the ground surface and more than 50% of the wheat seedlings in the field reached the standard time, it was the seedling emergence of the variety, and the seedling emergence period was the number of days from sowing to seedling emergence; the middle part of the young spike of the plant was exposed to the scabbard leaf sheaths as the standard of heading, and the heading period was the number of days from sowing to heading; the maturity period of wheat was the milky ripening period, and the standard was that the stalks and leaves were yellowish-green, and the kernels were milky with milky contents. The maturity period of wheat was recorded at the stage of milky ripening when the stalks and leaves were yellow-green, the kernels had milky inclusions, the kernels turned yellowish at the end of milky ripening, the water content of the stalks was 65%-75% and more than half of the accessions fulfilled the criterion, and the maturity period was the number of days from sowing to maturity. From seedling to heading stage (S1) and from seedling to maturity stage (S2) were calculated from the recorded data in days.

Methods of phenotypic analysis

The process of phenotyping was carried out by analysis of variance (ANOVA) as well as distributional evaluation of phenotypes, significance test of difference and correlation analysis, in addition all statistical analyses of data were implemented in IDE Spyder under Anacondas3 using Python 3.8.8 for data processing on a computer with an Intel i7-6800 K 3.40 GHz CPU, 16 GB of RAM, and an Nvidia GeForce GTX 2080Ti on a graphics workstation running the Win10 operating system. Statistical analyses, correlation analyses and tests of significance of differences were performed on wheat photosynthetic traits using the application software Pandas 1.3.2, Matplotlib 3.4.2, Scikit-Learn 0.24.2 and SPSS 21.0.

Chip typing and population structure analysis

The kernel DNA was extracted using the SDS method (Wang et al., 2014). The DNA quality was assessed by 1.2% agarose gel electrophoresis, and the DNA concentration was measured with a NanoDrop^{TM}ND-2000 spectrophotometer (Wang et al., 2014), and 239 wheat materials were scanned using Affymetrix Axiom 55K array (Beijing Boao Jingdian Biotechnology Co., Beijing, China). Illumina's Genome Studio Software was used for the original SNP typing of the samples. Markers with a filtration deletion rate of more than 20% and a minimum allele frequency (MAF) of less than 5%. High-quality SNP markers were retained for subsequent analysis (Figure S1).

The 2000 SNP markers after screening by random selection were screened for SNP markers that required a gene frequency greater than 10% and were evenly distributed on each different chromosome. Population structure analysis was performed by Structure v2.3.4 software (Zhu et al., 2008). The software was set up with reference to previous studies, and the results of the population structure were first presented visually as structure analysis plots through Tassel 5.0. Then the principal component analysis and Neighbor-jointing (NJ) evolutionary tree (Pritchard et al., 2000; Breseghello and Sorrells, 2006) were estimated for the population structure through Tassel 5.0, and finally plotted through the Matplotlib package for Python.

Linkage disequilibrium calculation

The squared correlation coefficient between loci (r^2) was used as a parameter to measure the linkage disequilibrium between two polymorphic loci between populations. r^2 was mainly calculated using Tassel 5.0 software and the 95th percentile of the r^2 value was used as a threshold to estimate the LD decay distance. The r^2 values between chained clusters were considered for square root transformation to account for the effect of the background of chained imbalances between chained clusters. Parameter values greater than the 95% of this distribution were used as thresholds to intercept the LD decay distance within the same chained cluster. Therefore, during chain analysis, the physical distance between loci was compared and when the physical distance was less than that LD decay distance, the loci were considered to be the same locus (Yang et al., 2007).

Association mapping

Association analysis, also known as linkage mapping or linkage disequilibrium mapping (LD mapping), is a quantitative genetic analysis technique that identifies loci or markers associated with a target trait based on the LD between alleles in different locis in a natural population by linking the diversity of the target trait to polymorphisms in the gene or marker.

In this study, we used Tassel v5.0 software combined with data for reproductive trait test and 55K SNP microarray data to carry out under different several reproductive stages and different water and drought treatments in wheat, and selected mixed linear model (MLM) for association analysis of the population. Through the analysis of the results calculated by the software, it is easy to see that the SNP marker can be identified as significantly associated with the trait when $P \leq 0.001$. Manhattan and QQ plots can be drawn from the results. The QQ plot can be used to further judge the correctness of the results of the association analysis and to exclude the appearance of some false positives. The horizontal coordinate of the Manhattan plot is the 21 chromosomes of wheat, and the vertical coordinate is the negative logarithm of the P-value of SNP markers. The distribution of SNP markers in all chromosomes and the loci of significant association can be seen through the Manhattan plot (Maccaferri et al., 2016).

Candidate gene prediction

The extended sequences of the stable SNP markers were subjected to BLAST comparison in the common wheat Chinese Spring genome database (https://urgi.versailles.inrae.fr/blast_iwgsc/) and gene function annotation in the Wheat Omics 1.0 database (http://wheatomics.sdau.edu.cn/) for gene function annotation.

Results

Analyses of phenotypic variation during the reproductive period

Two traits at S1 and S2 were obtained for phenotyping at four environments, E1, E2, E3 and E4, respectively. The data were evaluated through four dimensions, which are mean expressed as μ, median expressed as median, coefficient of variation expressed as cv (coefficient of variation), and standard deviation expressed as σ. From Figure 1, it can be seen that the trait μ of S1 stage in E1 environment was 194.82, median was 194.50, cv was 1.30% and σ was 2.57; the trait μ of S2 stage in E1 environment was 237.79, median was 237.00, cv was 1.50% and σ was 3.49; the trait of S1 stage in E2 environment μ was 192.65, median was 192.50, cv was 2.00%, and σ was 3.78; S2 trait μ was 241.60, median was 242.25, cv was 1.50%, and σ was 3.49 in E2 environment; and S1 trait μ was 225.26 in E3 environment, median was 224.50 with a cv of 1.00% and a σ of 2.37; the S2 trait μ in the E3 environment was 274.80 with a median of 275.00, a cv of 0.50%, and a σ of 1.27; the S1 trait μ in the E4 environment was 197.14 with a median of 197.00, a cv of 1.40%, and a σ of 2.75; the E4 environment the S2 trait μ was 236.81, median was 236.50, cv was 1.00%, and σ was 2.29. Overall the S1 trait μ was 192.65-225.26, median was 192.50-224.50, cv was 1.00%-2.00%, and σ was 2.37-3.78; the S2 trait μ was 236.81-274.80, median was 236.50-275.00, cv was 0.50%-1.50%, and σ was 1.27-3.49. The data in different environments showed a continuous and normal distribution, which is in line with typical quantitative trait characteristics.

By further analyzing the variance of the S1 and S2 traits of winter wheat obtained from different environments, Table 1 shows that the standard deviation ranges from 1.27-3.78, and analyses from the point of view of the analysis of variance indicate that the S1 and S2 traits of wheat are mainly genotypically determined, and also affected by the environment, and that the genetic factor is the main reason for its phenotypic variability.

Correlation analyses of reproductive traits (time to flowering and maturity)

The results of correlation analysis of two traits at S1 and S2 stage at four environmental points, E1, E2, E3 and E4, are shown in Figure 2. The correlations of reproductive traits of winter wheat in different environments were compared in the figure, from which it can be seen that most of the winter wheat reproductive traits reached the highly significant level ($p<0.001$).The correlations of S1 stage traits in E1 environment were -0.02-0.67, and the correlations of S2 stage traits in E1 environment were 0.22-0.55; and the correlations of S1 stage traits in E2 environment were 0.37-0.55; and the correlations of S2 stage traits in E3 and E4 environments were 0.37-0.55. correlation was 0.37-0.63 for S1 stage trait in E2 environment and 0.45-0.62 for S2 stage trait in E2 environment; correlation of S1 stage trait in E3 environment was 0.41-0.62 and correlation of S2 stage trait in E4 environment was 0.36-0.62. Correlation of S1 stage trait in different environments was 0.45-0.63 and correlation of S2 stage trait in different environments was 0.45-0.63. Overall the correlation of S1 and S2 of bread wheat in different environments from 0.13-0.81, reaching highly significant levels ($p<0.001$).

Population structure and evolutionary tree analysis

Using Structure software to analyze the population structure of 239 test materials (Figure 3), the group structure was divided by group structure, evolutionary tree and principal component analysis. The results showed that the results of the three analysis methods were consistent, and it was reasonable to divide the whole population into three subgroups, of which subgroup 1 had 95 accessions (lines), subgroup 2 had 89 accessions (lines), and subgroup 3 had 55 accessions (lines). The distribution frequencies of the materials contained in the three subgroups were 39.75%, 37.24% and 23.01% in the following order. The LD decay distances of 239 wheat accessions (lines) in genomes A, B, D and the whole genome were calculated to be 3, 3, 2 and 3 Mb, respectively. Based on the LD decay distances of the whole genome, the loci within the interval of 3 Mb before and after the physical map were identified as a candidate locus.

GWAS analysis of reproductive traits (time to flowering and maturity)

The S1 and S2 of 239 wheat accessions (lines) were combined with 16,649 high-quality SNP markers typed by 55K SNP chip

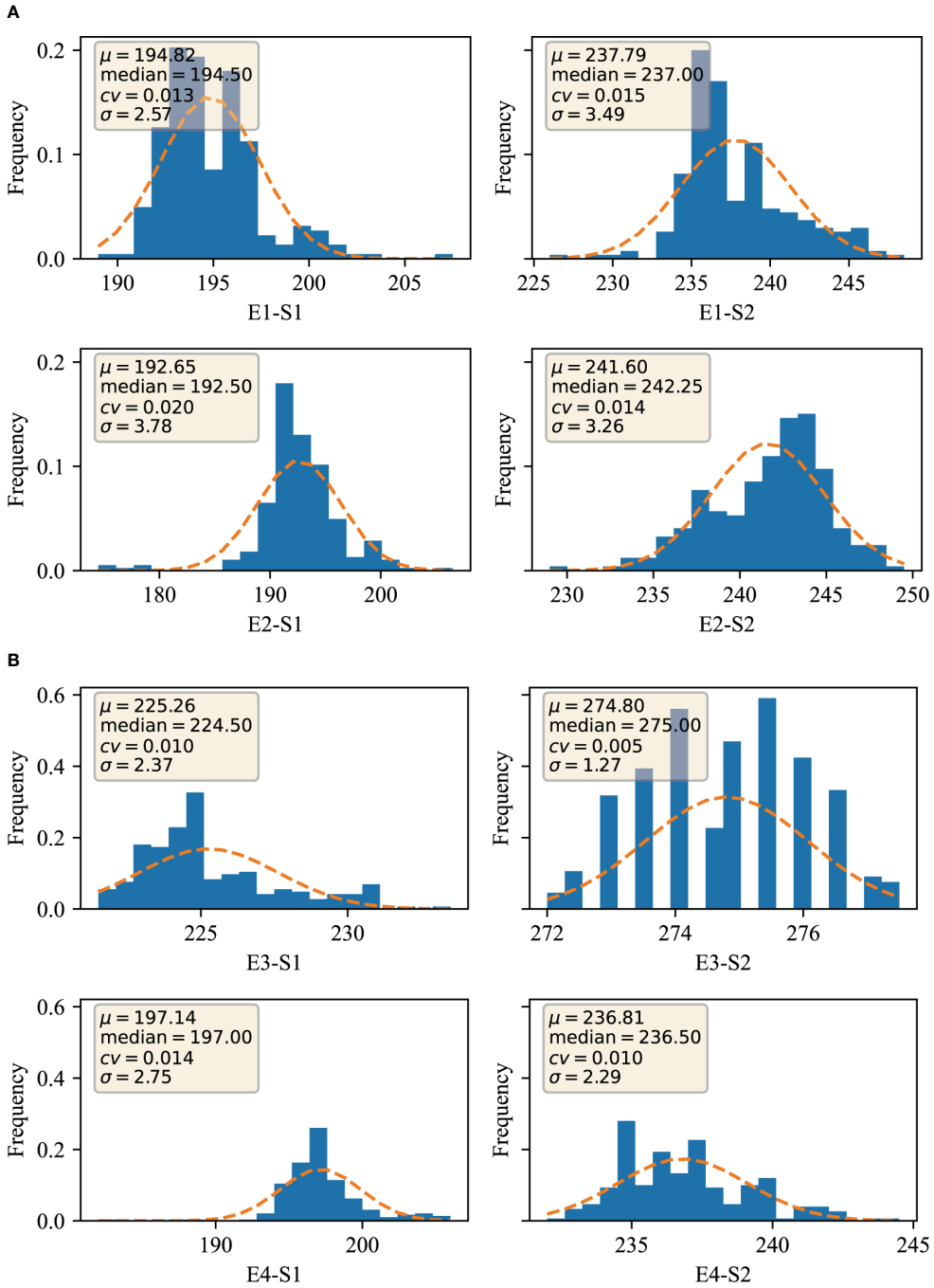


FIGURE 1
Distribution of traits in different environments **(A)** Germination-spike and germination-maturity in E1 and E2 environments; **(B)** Germination-spike and germination-maturity in E3 and E4 environments. E1: 2020 Zepu, Xinjiang; E2: 2021 Zepu, Xinjiang E3: 2021 Anning Drain, Xinjiang E4: 2022 Zepu, Xinjiang; S1: heading stage; S2: maturity stage.

screening for genome-wide association analysis using TASSEL 5.0 software. Based on the MLM (Q+K) model, the markers were considered to be significantly associated with the trait when $P \leq 0.001$, and loci detected in multiple environments were considered to be stably heritable (Figure 4, Table 2, Table S2). Analysis of the GWAS results showed that a total of 238 SNP markers were detected for the S1 trait, of which a total of eight markers were detected in multiple environments, distributed on chromosomes 1B、2D、3A、5B、6D、7A、7D with individual interpretable phenotypic variation rates of 4.03%-16.06%. *AX-109375483* and *AX-110425403* located on chromosome 1B, were both detected simultaneously in both environments, with phenotypic variation rates of 4.49%-4.96% and 4.82%-16.06%, while *AX-108940388* located on chromosome 2D, was detected simultaneously in both environments E2 and E4, with phenotypic variation rates of 5.03%-15.08%; *AX-110591324* located on chromosome 3A was detected in both E2 and E4 environments at the same time with a phenotypic variation rate of 5.16%-15.31%; *AX-109429484* located on

10.3389/fpls.2023.1296197

TABLE 1 Descriptive statistics of wheat reproductive traits (time to flowering and maturity).

Environments	Traits	Max	Min	Mean	Standard Deviation	Cv (%)
E1	S1	207.5	189	194.83	2.57	1.32
	S2	248.5	226	237.79	3.5	1.47
E2	S1	206.5	174.5	192.65	3.78	1.96
	S2	249.5	229	241.6	3.27	1.35
E3	S1	233.5	221.5	225.26	2.37	1.05
	S2	277.5	272	274.8	1.27	0.46
E4	S1	206	182	197.14	2.75	1.4
	S2	244.5	232	236.81	2.3	0.97

chromosome 5B was detected in both E2 and E4 environments at the same time with a phenotypic variation rate of 4.03%-15.53%; *AX-111919223* located on chromosome 6D, was detected in both E2 and E4, with a phenotypic variation rate of 5.20%-6.73%; *AX-110961085* located on chromosome 7A, was detected in both E2 and E4, with a phenotypic variation rate of 4.14%-8.09%; and *AX-108866484* located on chromosome 7D, was detected in both E2 and E4, with a phenotypic variation rate of 4.14%-8.09%. *AX-108866484* located on chromosome 7D was detected in both E3 and E4 environments, with a phenotypic variation rate of 4.36%-15.91%, respectively. A total of 55 SNP markers were detected for the S2 trait, of which *AX-110986688* located on chromosome 1A was detected in both E1 and E2 environments, with a single locus explaining the phenotypic variation rate of 4.18%-5.08%.

Functional prediction of candidate genes for reproductive traits (time to flowering and maturity)

SNP markers with large phenotypic effect values that could be stably inherited were searched in the Chinese Spring Genome

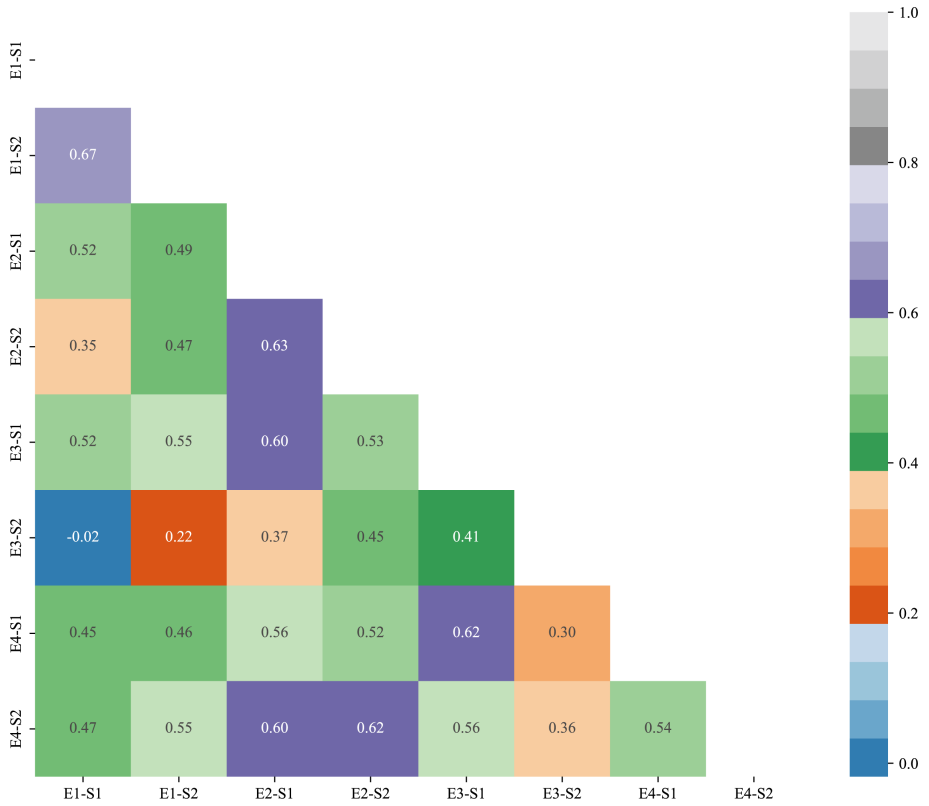


FIGURE 2
Correlation plots of germination-sprouting and germination-maturity traits in different environments.

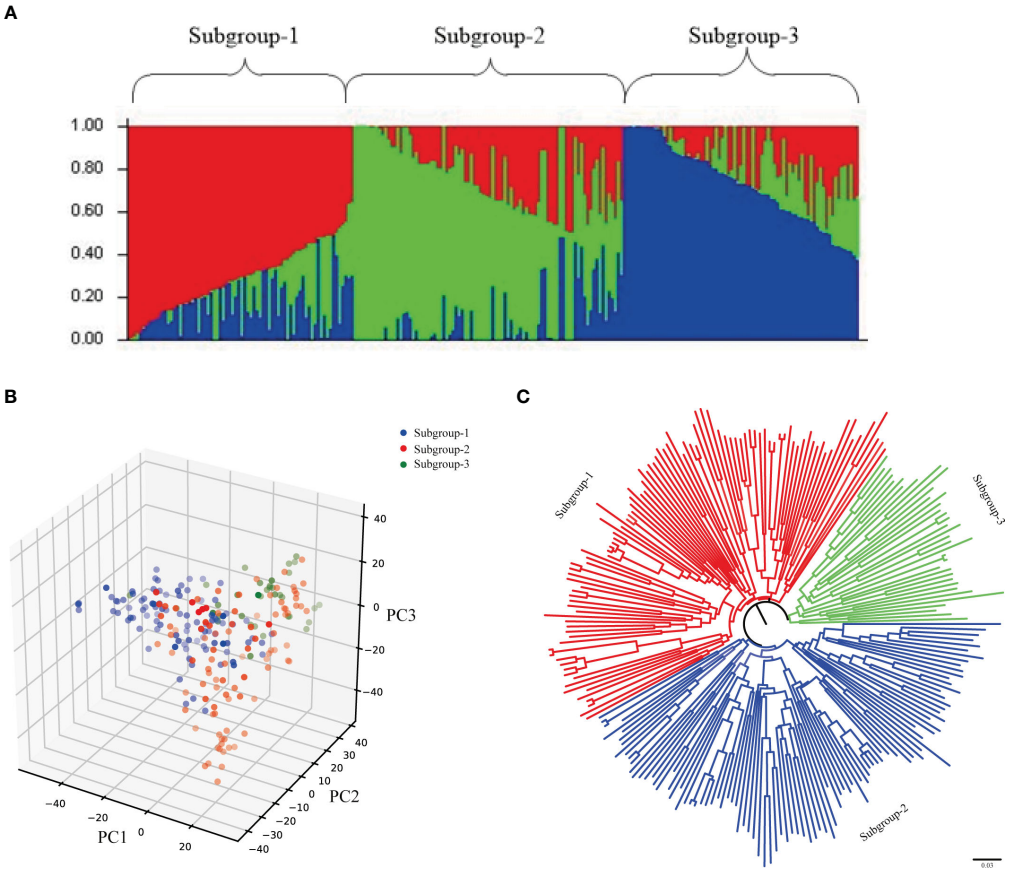


FIGURE 3
Population structure analysis of 239 wheat accessions. **(A)** Population structure analysis; **(B)** Neighbor-joining method evolutionary tree; **(C)** Principal component analysis.

Database of common wheat and BLASTx sequence comparison was performed in the NCBI database, and a total of nine candidate genes most likely to be associated with reproductive traits were mined (Table 3).

Candidate genes for reproductive traits (time to flowering and maturity) are mainly associated with photosynthesis, Ca^{2+} transport, phytohormone biosynthesis and signal transduction in crops. The genes *TraesCS1B01G312100* and *TraesCS1B01G356000* located on chromosome 1B encode glycosyltransferases and F-box family proteins, respectively; *TraesCS2D01G044700* located on chromosome 2D is associated with cytochrome proteins; *TraesCS3A01G036000* on chromosome 3A encodes a zinc finger family protein; *TraesCS3A01G03600* on chromosome 3A encodes a zinc finger family protein; and *TraesCS3A01G036000* encodes a zinc finger family protein; *TraesCS5B01G288700*, located on chromosome 5B, encodes an S-acyltransferase; *TraesCS6D01G404800*, located on chromosome 6D, encodes a calcium-dependent protein kinase; *TraesCS7A01G404800* located on chromosome 7A, encodes a calcium-dependent protein kinase; and *TraesCS7A01G560200* encodes Photosystem II stability/assembly factor HCF136; *TraesCS7D01G098100* on chromosome 7D encodes zinc finger protein; *TraesCS1A01G362500* on chromosome 1A encodes Cytokinin riboside 5'-monophosphate phosphoribohydrolase.

Discussion

Time to heading and maturity trait association analyses

With the rapid development of biology and bioinformatics, GWAS analysis has become an important way to study quantitative traits in plants, and the mining of genes related to reproductive traits (time to flowering and maturity) in wheat has been promoted to a greater extent. Meanwhile, reproductive traits are typical quantitatively inherited traits, which are subject to the joint action of genotype and environment. In this study, through four environmental sites and three years of data accumulation, a total of nine stable genetic loci in multiple environments were excavated to be significantly or very significantly associated with wheat heading and maturity stages, which affect wheat adaptation and yield and are complex quantitative traits regulated by multiple genes. It has been shown that the genes associated with heading and maturity stages are mainly located on chromosomes 1A、2A、 2B、2D、3A、5A、5B、5D and 7B (Pritchard et al., 2000; Somers et al., 2004; Evanno et al., 2005; Liu and Muse, 2005; Jingna et al., 2014; Shi et al., 2019). Some SNP loci on chromosomes 1B、3D and 7D were identified by GWAS analysis as significantly associated with

10.3389/fpls.2023.1296197

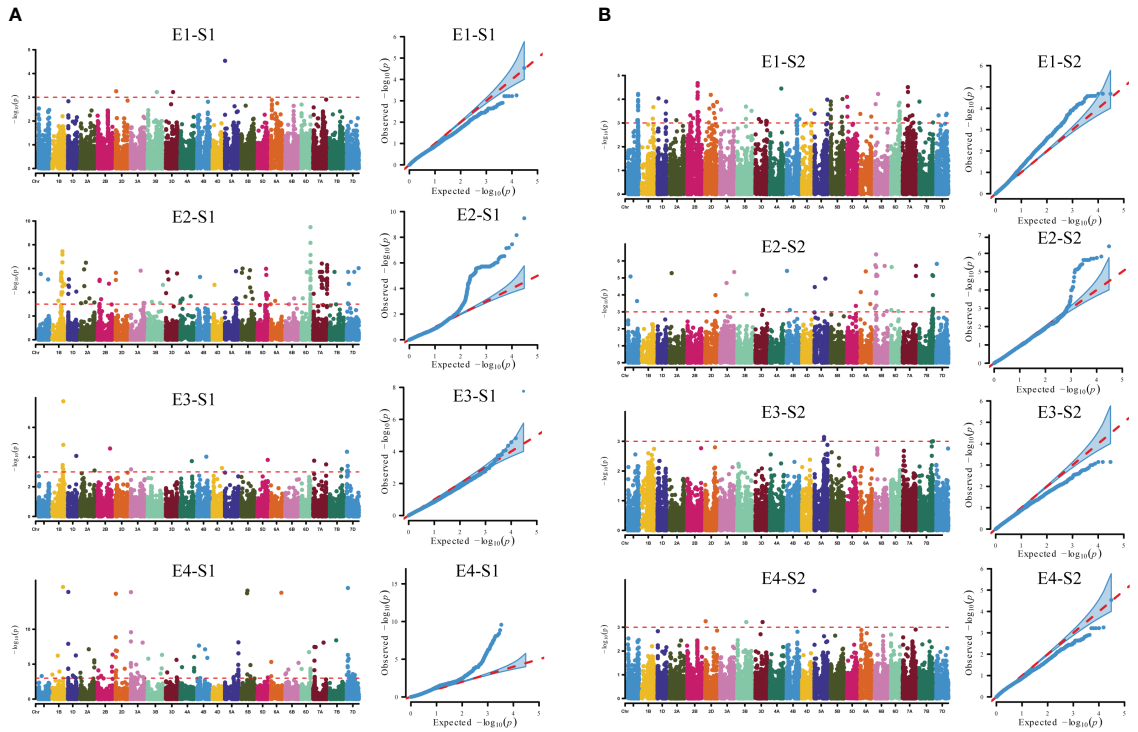


FIGURE 4
GWAS of reproductive traits in different environments. **(A)** Manhattan and quantile-quantile plots for S1 in different environments. **(B)** Manhattan and quantile-quantile plots for S2 in different environments. E1: 2020 Zepu, Xinjiang; E2: 2021 Zepu, Xinjiang E3: 2021 Anning Drain, Xinjiang E4: 2022 Zepu, Xinjiang; S1: heading stage; S2: maturity stage.

heading (Hardy and Vekemans, 2002; Zou et al., 2017; Ma et al., 2021; Zhang et al., 2021). In this study, we identified that the SNPs significantly associated with heading were mainly located on chromosomes 1A、1B、2D、3A、5B、6D and7A, and the results were more consistent with the previous localization results.

Functional analysis of candidate genes

The GWAS was used to detect 9 SNP marker locis that were significantly associated with reproductive traits (time to flowering and maturity) in wheat, and nine candidate genes that might be related to reproductive traits were screened in the Chinese Spring Genome Database of common wheat. The genes *TraesCS1B01G312100* and *TraesCS1B01G356000* located on 1B encode glycosyltransferases and F-box family proteins, respectively; glycosylation is an important post-translational modification of proteins in plants, which is involved in the regulation of various biological functions. Glycosyl transferase (GT) is one of the most important enzymes in the class of glycosylation enzymes. F-box family proteins have important roles in physiological processes such as phytohormone signaling, light signaling and floral organ development (Bi et al., 2006). The gene

TABLE 2 Information of significantly associated loci of wheat reproductive traits (time to flowering and maturity).

Traits	Marker	Chr.	Position (Mb)	P-value	R^2 (%)	Environment
S1	*AX-109375483*	1B	535.23-536.57	1.10E-05-3.21E-05	4.49-4.96	E2、E3
	AX-110425403	1B	581.10-587.17	8.792E-17-1.50E-05	4.82-16.06	E3、E4
	AX-108940388	2D	12.92-17.45	8.38E-16-9.26E-06	5.03-15.08	E2、E4
	AX-110591324	3A	16.63-20.15	4.86E-16-6.82E-06	5.16-15.31	E2、E4
	AX-109429484	5B	466.61-474.97	2.97E-16-9.37E-05	4.03-15.53	E2、E4
	AX-111919223	6D	466.89-472.91	1.87E-07-9.68E-06	5.20-6.73	E2、E4
	AX-110961085	7A	730.53-735.46	8.14E-09-3.12E-06	4.14-8.09	E2、E4
	AX-108866484	7D	57.69-59.71	1.24E-16-4.36E-05	4.36-15.91	E3、E4
S2	*AX-110986688*	1A	542.72-543.95	8.24E-06-6.61E-05	4.18-5.08	E1、E2

TABLE 3 Information of candidate genes for wheat reproductive traits (time to flowering and maturity).

Maker	Chr	Physical location (Mb)	Genes	Gene annotation
AX-109375483	1B	535.23-536.57	*TraesCS1B01G312100*	Glycosyltransferase
AX-110425403	1B	581.10-587.17	*TraesCS1B01G356000*	F-box protein
AX-108940388	2D	12.92-17.45	*TraesCS2D01G044700*	Cytochrome P450
AX-110591324	3A	16.63-20.15	*TraesCS3A01G036000*	Zinc finger family protein
AX-109429484	5B	466.61-474.97	*TraesCS5B01G288700*	S-acyltransferase
AX-111919223	6D	466.89-472.91	*TraesCS6D01G404800*	Calcium-dependent protein kinase
AX-110961085	7A	730.53-735.46	*TraesCS7A01G560200*	Photosystem II stability/assembly factor HCF136
AX-108866484	7D	57.69-59.71	*TraesCS7D01G098100*	Zinc transporter
AX-110986688	1A	542.72-543.95	*TraesCS1A01G362500*	Cytokinin riboside 5'-monophosphate phosphoribohydrolase

TraesCS2D01G044700, located on 2D, is related to cytochrome proteins; cytochrome P450 is an important oxidase in the microsomal mixed-function oxidase family, which is widely distributed in living organisms, and is involved in the synthesis and metabolism of a wide range of endogenous and exogenous compounds, which have important functions in biological oxidation, nitrogen fixation, photosynthesis, energy conversion, and storage (Himi et al., 2011). *TraesCS3A01G036000* and *TraesCS7D01G098100* located on chromosomes 3A and 7D encode zinc finger family proteins; which can be involved in physiological and biochemical regulatory mechanisms in plants during growth and development. *TraesCS5B01G288700* located on chromosome 5B encodes S-acyltransferase, which is a metabolite in plants that has an important role in growth and development and under drought stress. The gene *TraesCS6D01G404800* located on 6D encodes calcium-dependent protein kinase; this enzyme acts as a cellular second messenger, Ca^{2+} coordinates the perception of various physiological responses in plants, and the Ca^{2+} sensor transmits calcium signals downstream and triggers a cascade of reactions, regulating the processes of plant growth, development and response to the environment. The gene *TraesCS7A01G560200* on chromosome 7A encodes Photosystem II stability/assembly factor HCF136; PS II is a pigmented protein complex present in the membranes of cysts in plants that drives the light-activated transfer of electrons from water to plastocysts, accompanied by the production of molecular oxygen. The gene *TraesCS1A01G362500* located on chromosome 1A encodes Cytokinin riboside 5'-monophosphate phosphoribohydrolase. Cytokinins are a class of N6-adenine analogues, which are closely related to crop yield and plays a key role in the regulation of plant growth and development, including the promotion of fruiting, the release of apical dominance, the promotion of cell division, and the shortening of the transition from nutrient growth to reproductive growth. The analysis of gene function lays the foundation for our next step of functional marker development.

Conclusion

In this study, a comprehensive genome-wide association analysis was conducted on 239 wheat accessions (lines) from both domestic and international sources. The analysis focused on the heading and maturity stages of wheat using a 55K SNP microarray and a Q+K mixed linear model. The aim was to identify SNP markers associated with these important developmental phases.

The analysis identified a total of 293 SNP marker loci that showed significant associations ($P \leq 0.001$) with heading and maturity stages in wheat. Among these markers, nine were found to be consistently associated with these traits in multiple environments, indicating their stability and reliability. These stable SNP marker loci were located on different chromosomes of wheat, including 1A、1B、2D、3A、5B、6D and 7A. Furthermore, the researchers investigated the phenotypic effect values and inheritance patterns associated with these SNP markers. By searching the Chinese Spring Genome Database of common wheat, we identified nine candidate genes that were most likely to be associated with the heading and maturity stages of wheat. These candidate genes were selected based on their significant phenotypic effect values and stable inheritance patterns. It is important to note that the specific details of the candidate genes were not provided in the information provided. However, the identification of these candidate genes suggests their potential involvement in regulating the heading and maturity phases of wheat development.

Data availability statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Author contributions

YD: Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing. HF: Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing. YG: Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing.

GF: Formal Analysis, Writing – original draft, Writing – review & editing. XS: Formal Analysis, Writing – original draft, Writing – review & editing. SY: Data curation, Investigation, Methodology, Writing – review & editing. SD: Data curation, Investigation, Methodology, Writing – review & editing. TH: Data curation, Investigation, Methodology, Writing – review & editing. WW: Data curation, Investigation, Methodology, Writing – review & editing. JS: Data curation, Investigation, Methodology, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The present study was funded by the Xinjiang Academy of Agricultural Sciences Youth Science and Technology Backbone Innovation Ability Training Project (xjnkq-2020005); Xinjiang Academy of Agricultural Sciences Youth Science and Technology Backbone Innovation Ability Training Project (xjnkq-2021006); Xinjiang Uygur Autonomous Region Department of Agriculture and Rural Affairs 2022 "Unveiling and Hanging the Marshal" Research Project; Major Science and Technology Special Project of Xinjiang Autonomous Region (2021A02001-1).

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.2023.1296197/full#supplementary-material

SUPPLEMENTARY FIGURE 1
Density distribution of SNPs on chromosomes.

References

Bi, C. L., Liu, X., and Zhang, X. Y. (2006). The function of F-box protein in plant growth and development. *Hereditas* 28 (10), 1337–1342. doi: 10.3321/j.issn:0253-9772.2006.10.026

Breseghello, F., and Sorrells, M. E. (2006). Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172 (2), 1165–1177. doi: 10.1534/genetics.105.044586

Chen, Z., Cheng, X., Chai, L., Wang, Z., Du, D., Wang, Z., et al. (2020). Pleiotropic QTL influencing spikelet number and heading date in common wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 133, 1825–1838. doi: 10.1007/s00122-020-03556-6

Cheng, Y., Jiang, J., Chen, Q., Wang, Z., Zeng, M., Qin, F., et al. (2023). Radio-frequency treatment of medium-gluten wheat: effects of tempering moisture and treatment time on wheat quality. *J. Sci. Food Agric.* 103 (9), 4441–4449. doi: 10.1002/jsfa.12539

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

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2 (4), 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Himi, E., Maekawa, M., Miura, H., and Noda, K. (2011). Development of PCR markers for Tamyb10 related to R-1, red grain color gene in wheat. *Theor. Appl. Genet.* 122, 1561–1576. doi: 10.1007/s00122-011-1555-2

Hoogendoorn, J. (1985). A reciprocal F_1 monosomic analysis of the genetic control of time of ear emergence, number of leaves and number of spikelets in wheat (*Triticum aestivum* L.). *Euphytica* 34 (2), 545–558. doi: 10.1007/bf00022954

Jingna, R., Yang, Y., and Fanfan, D. (2014). Analysis of QTL for heading date and interaction effects with environments in wheat. *J. Triticeae Crops* 34 (9), 1185–1190. doi: 10.7606/j.issn.1009-1041.2014.09.04

Liu, K., and Muse, S. V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* 21 (9), 2128–2129. doi: 10.1093/bioinformatics/bti282

Ma, S., Wang, M., Wu, J., Guo, W., Chen, Y., Li, G., et al. (2021). WheatOmics: A platform combining multiple omics data to accelerate functional genomics studies in wheat. *Mol. Plant* 14 (12), 1965–1968. doi: 10.1016/j.molp.2021.10.006

Maccaferri, M., El-Feki, W., Nazemi, G., Salvi, S., Canè, M. A., Colalongo, M. C., et al. (2016). Prioritizing quantitative trait loci for root system architecture in tetraploid wheat. *J. Exp. Bot.* 67 (4), 1161–1178. doi: 10.1093/jxb/erw039

Mackay, I., and Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.* 12 (2), 57–63. doi: 10.1016/j.tplants.2006.12.001

Morales, F., Ancín, M., Fakhet, D., González-Torralba, J., Gámez, A. L., Seminario, A., et al. (2020). Photosynthetic metabolism under stressful growth conditions as a bases for crop breeding and yield improvement. *Plants* 9 (1), 88. doi: 10.3390/plants9010088

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *Am. J. Hum. Genet.* 67 (1), 170–181. doi: 10.1086/302959

Shi, C., Zhao, L., Zhang, X., Lv, G., Pan, Y., and Chen, F. (2019). Gene regulatory network and abundant genetic variation play critical roles in heading stage of polyploidy wheat. *BMC Plant Biol.* 19 (1), 1–16. doi: 10.1186/s12870-018-1591-z

Somers, D. J., Isaac, P., and Edwards, K. (2004). A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 109, 1105–1114. doi: 10.1007/s00122-004-1740-7

Song, Y. X., Jing, R. L., Huo, N. X., Ren, Z. L., and Jia, J. Z. (2006). Detection of QTLs for heading in common wheat (*T.aestivum* L.) using different populations. *Scientia Agricultura Sin.* 11), 2186–2193. doi: 10.3321/j.issn:0578-1752.2006.11.004

Sourdille, P., Snape, J., Cadalen, T., Charmet, G., Nakata, N., Bernard, S., et al. (2000). Detection of QTLs for heading time and photoperiod response in wheat using a doubled-haploid population. *Genome* 43 (3), 487–494. doi: 10.1139/g00-013

Wang, K. S., Dong, S. S., Li, F. J., Guo, J., Tai, S. Q., Wang, L. B., et al. (2020). QTL mapping and analysis of heading time and flowering time of wheat. *Shandong Agric. Sci.* 52 (01), 17–23. doi: 10.14083/j.issn.1001-4942.2020.01.003

Wang, Y., Fan, Q. Q., Zhang, L., Sui, X. X., Li, G. Y., Chu, X. S., et al. (2007). Genetic analysis on precocialism of wheat variety K35. *J. Triticeae Crops* 27 (6), 957–960. doi: 10.3969/j.issn.1009-1041.2007.06.003

Wang, S., Zhu, Y., Zhang, H., Chang, C., and Ma, C. (2014). Analysis of genetic diversity and relationship among wheat breeding parents by SSR markers. *J. Triticeae Crops* 34 (5), 621–627. doi: 10.7606/j.issn.1009-1041.2014.05.08

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

Yan, X., Shi, Y. G., Li, X. Y., Wang, S. G., and Sun, D. Z. (2015). QTL mapping for flowering time in wheat. *J. Shanxi Agric. Sci.* 43 (8), 919–921. doi: 10.3969/j.issn.1002-2481.2015.08.01

Yang, Y., Zhao, X., Xia, L., Chen, X., Xia, X., Yu, Z., et al. (2007). Development and validation of a Viviparous-1 STS marker for pre-harvest sprouting tolerance in Chinese wheats. *Theor. Appl. Genet.* 115, 971–980. doi: 10.1007/s00122-007-0624-z

Zhang, L., Zhang, H., Qiao, L., Miao, L., Yan, D., Liu, P., et al. (2021). Wheat MADS-box gene *TaSEP3-D1* negatively regulates heading date. *Crop J.* 9 (5), 1115–1123. doi: 10.1016/j.cj.2020.12.007

Zhu, C., Gore, M., Buckler, E. S., and Yu, J. (2008). Status and prospects of association mapping in plants. *Plant Genome* 1 (1), 5. doi: 10.3835/plantgenome2008.02.0089

Zou, J., Semagn, K., Chen, H., Iqbal, M., Asif, M., N'Diaye, A., et al. (2017). Mapping of QTLs associated with resistance to common bunt, tan spot, leaf rust, and stripe rust in a spring wheat population. *Mol. Breed.* 37, 1–14. doi: 10.1007/s11032-017-0746-1

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 04 December 2023
DOI 10.3389/fpls.2023.1294507

OPEN ACCESS

EDITED BY
Patricio Hinrichsen,
Agricultural Research Institute, Chile

REVIEWED BY
Erika Salazar,
Investigaciones Agropecuarias de Chile (INIA), Chile
Rodrigo Iván Contreras-Soto,
Universidad de O'Higgins, Chile

*CORRESPONDENCE
Thomas Lübberstedt
thomasl@iastate.edu

RECEIVED 14 September 2023
ACCEPTED 17 November 2023
PUBLISHED 04 December 2023

CITATION
Ledesma A, Santana AS, Sales Ribeiro FA, Aguilar FS, Edwards J, Frei U and Lübberstedt T (2023) Genome-wide association analysis of plant architecture traits using doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic maize population. *Front. Plant Sci.* 14:1294507. doi: 10.3389/fpls.2023.1294507

Genome-wide association analysis of plant architecture traits using doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic maize population

Alejandro Ledesma[1], Alice Silva Santana[2], Fernando Augusto Sales Ribeiro[3], Fernando S. Aguilar[4], Jode Edwards[5], Ursula Frei[6] and Thomas Lübberstedt[6]*

[1]National Institute of Forestry, Crop and Livestock Research, Tepatitlán, Jalisco, Mexico, [2]Department of Agronomy, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, [3]Department of Agronomy, Federal University of Lavras, Lavras, Minas Gerais, Brazil, [4]Colombian Sugarcane Research Center (Cenicana), Cali, Cauca Valley, Colombia, [5]U.S. Department of Agriculture, Agricultural Research Service, Ames, IA, United States, [6]Department of Agronomy, Iowa State University, Ames, IA, United States

Selection in the Iowa Stiff Stalk Synthetic (BSSS) maize population for high yield, grain moisture, and root and stalk lodging has indirectly modified plant architecture traits that are important for adaptation to high plant density. In this study, we developed doubled haploid (DH) lines from the BSSS maize population in the earliest cycle of recurrent selection (BSSS), cycle 17 of reciprocal recurrent selection, [BSSS(R)17] and the cross between the two cycles [BSSS/BSSS(R)C17]. We aimed to determine the phenotypic variation and changes in agronomic traits that have occurred through the recurrent selection program in this population and to identify genes or regions in the genome associated with the plant architecture changes observed in the different cycles of selection. We conducted a *per se* evaluation of DH lines focusing on high heritability traits important for adaptation to high planting density and grain yield. Trends for reducing flowering time, anthesis-silking interval, ear height, and the number of primary tassel branches in BSSS(R)17 DH lines compared to BSSS and BSSS/BSSS(R)C17 DH lines were observed. Additionally, the BSSS(R)C17 DH lines showed more upright flag leaf angles. Using the entire panel of DH lines increased the number of SNP markers identified within candidate genes associated with plant architecture traits. The genomic regions identified for plant architecture traits in this study may help to elucidate the genetic basis of these traits and facilitate future work about marker-assisted selection or map-based cloning in maize breeding programs.

KEYWORDS

candidate gene, quantitative trait locus, diversity, genetic resources, *Zea mays*

1 Introduction

Genetic variability is essential in plant breeding programs. Plant breeders primarily focus on short-term breeding goals, because of the need to deliver new varieties. This may result in a narrow genetic base of maize elite germplasm (Andorf et al., 2019) and could lead to a yield plateau, increased vulnerability to pests and make it difficult to meet new market demands (Pollak, 2003). Assessment of the genetic variability that exists in available germplasm is fundamental for crop improvement. Genetic improvement of important agronomic traits while maintaining genetic variability long-term is desirable in maize breeding programs (Hallauer and Darrah, 1985). In this context, recurrent selection procedures in maize have proven to be effective to increase the frequency of superior lines for grain yield and other agronomic traits while maintaining genetic variability (Hallauer and Darrah, 1985; Ordas et al., 2012; Fritsche-Neto et al., 2023). Recurrent selection is the systematic selection of desirable individuals from a population followed by the selected individuals' recombination to form the next selection cycle. It was suggested by Jenkins (1940) as a method of intrapopulation improvement and later described for population improvement using a tester (Hull, 1945). The most significant advantage of this method is the increase in the population's mean performance for one or more traits by increasing the frequency of favorable alleles while maintaining genetic variability for continued genetic improvement. Genetic variability will be preserved if an adequate number of lines is intermated for the next selection cycle.

The Iowa Stiff Stalk Synthetic (BSSS) maize population (Sprague, 1946) has undergone recurrent selection since 1939. This population was developed by intermating 16 inbred lines selected by various maize breeders for superior stalk quality. Of these progenitors, 10 were derived from multiple strains of the Reid Yellow Dent open-pollinated population, 4 had miscellaneous origins, and the genetic background of 2 is unknown (Sprague, 1946). Two recurrent selection programs were initiated in BSSS, a half-sib program with the double cross hybrid IA13 used as a tester and a reciprocal program with the Iowa Corn Borer Synthetic Number One (BSCB1) (Penny and Eberhart, 1971; Eberhart et al., 1973; Martin and Hallauer, 1980; Smith, 1983; Helms et al., 1989; Keeratinijakal and Lamkey, 1993). Two additional programs were initiated using the population generated by seven cycles of half-sib selection (Lamkey, 1992; Edwards 2010). In all four programs selection was carried out for increased grain yield, low grain moisture at harvest, and decreased root and stalk lodging. Several important inbred lines have been developed from the BSSS population (B14, B37, B73, and B84). They have made significant contributions to the maize industry in the US, especially B73, one of the most successful maize inbred lines developed in the public sector and benefited industry and farmers substantially (Coffman et al., 2019).

Agronomic and plant architecture changes have been reported for different selection cycles in the BSSS maize population. These changes involve modifications in traits such as plant height, anthesis-silking interval, leaf angle and number of tassel branches (Brekke et al., 2011; Edwards, 2011). Changes in plant architecture traits over continuous selection cycles, driven by testing under higher population densities have increased throughout the hybrid era (Brekke et al., 2011). In response to increasing plant densities over time, genotypes from later cycles of recurrent selection should have more upright leaves, reduced anthesis-silking interval, and fewer tassel branches.

Genome-wide association studies (GWAS) are a useful tool for analyzing allelic diversity to identify superior alleles and dissect the genetic architecture, which furthers genetic improvement in crops. The increasing application of association mapping is due to the rapid development of sequencing and DNA marker techniques, which resulted in cost-effective high-throughput genotyping technologies. Genomic regions and candidate genes conferring adaptation to high plant density identified by GWAS could help to speed up genetic resource utilization. Identifying genomic regions associated with plant architecture changes may help to unlock genetic resources not adapted to high plant densities, either by selecting such regions in genetic resource populations like early cycles of recurrent selection programs or after introducing them into respective materials (Zhao et al., 2019).

In this study, we phenotypically characterized DH lines developed from the unselected base population, BSSS, the 17th cycle of reciprocal recurrent selection BSSS(R)C17, and the cross between them BSSS/BSSS(R)C17. The purpose of this study was to i) investigate changes in phenotypic diversity for plant architecture traits among DH lines developed from the earliest and the most advanced selection cycle, ii) identify DH lines with both significant C0 background and modern plant architecture traits conferring adaptation to high plant density that could be used as genetic resources, iii) evaluate how to best use DH lines for GWAS from the two subpopulations BSSS and BSSS(R)C17 and the cross between them, to identify regions affecting plant architecture traits, and iv) determine the inheritance of those regions, in particular, whether major genes are involved that may help to accelerate recurrent selection cycles to adapt any germplasm to modern plant types.

2 Materials and methods

2.1 Breeding populations

Two synthetic populations BSSS, BSSS(R)C17, and the cross between them, BSSS/BSSS(R)C17, representing different stages of cycle advancement in the recurrent selection program of the Iowa Stiff Stalk Synthetic maize population, BSSS, were used to develop DH lines. The synthetic BSSS corresponds to the unselected base population (C0) formed by intermating 16 inbred lines selected for above average stalk quality in 1934 (Sprague, 1946). The C0 seed used came from subsequent cycles of seed multiplication in C0 for maintenance over time. The BSSS(R)C17 population corresponds to the most advanced cycle (C17) available when research was initiated to study crosses between the unselected base population and an advanced cycle. The cross BSSS/BSSS(R)C17 was created by intermating plants from BSSS and BSSS(R)C17.

2.2 Doubled haploid line development

Random samples (~1400 plants) from BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17 were pollinated with BHI301, a maternal haploid inducer (Almeida et al., 2020), in an isolation field to generate the haploid seed. Seeds produced from these plants expressing the *R-nj* marker gene in the endosperm but not in the embryo were classified as haploid. The haploid seed was then germinated in plug trays in a greenhouse at the Department of Agronomy, Iowa State University (ISU). Once seedlings developed 2-3 leaves, a colchicine treatment was applied following the DH Facility protocol at ISU (Vanous et al., 2017). Two days after the colchicine treatment, haploid seedlings were transplanted in the field at the Agricultural Engineering and Agronomy Research Farm, Boone, Iowa. Putative DH0 plants shedding pollen were self-pollinated to produce DH1 generation seed. Seed multiplication was performed during subsequent growing seasons, and lines were screened for uniformity and discarded if segregating. In total, 135, 194 and 187 DH lines from BSSS (C0_DHL), BSSS(R)17 (C17_ DHL) and BSSS/BSSS(R)C17 (C0/C17_DHL), respectively, were obtained.

2.3 Experimental design and phenotypic data collection

The 516 DH lines plus 16 progenitors of the BSSS population [A3G-3-3-1-3, CI 540, Fe (Parent of F1B1), I-159, IL12E, B2 (Parent of F1B1), Oh 3167B, Os 420, Tr 9-1-1-6, WD 456, I224, LE23, Ind. 461, Hy, AH83, CI 187-2] and the inbred line B73 were planted during summer 2019 at three locations: Plant Introduction Station (PI) in Ames, IA, Johnson Farm near Kelly, IA, and Burkey at Agronomy Farm near Boone, IA. The experiment was planted in each location using a modified split plot design with two replications, where the DH lines for populations BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17 constituted the whole plot treatment factor and the DH lines within each population the subplot treatment factor. This design differs from a classical split-plot because the subplot factor (DH lines) was nested within the whole-plot factor, population. Progenitors were included as subplot treatments within BSSS whole plots. Inbred line B73 was used as a check and replicated 14 times within each replicate resulting in 546 experimental units per replication (516 DH lines, 16 progenitors, and 14 replicates of B73) The subplot experimental unit consisted of a single row plot, 3.8 m long with 15 plants with 0.76 m between rows. The whole-plot factor experimental unit was a block containing 39 subplots arranged side by side. Each replication, containing 546 subplots, was divided into three whole plots, which were then separated into 4, 5, and 5 blocks for C0_DHL, C17_DHL, and C0C17_DHL, respectively. Each whole-plot block was randomly assigned to a range in the field.

Phenotypic data were collected on a plot basis for male flowering, female flowering, plant height, ear height, flag leaf angle, tassel length, and the number of primary tassel branches. Male flowering and female flowering were recorded as the date when 50% of the plants in the row were shedding pollen and had visible silks, respectively. Plants were recorded as shedding pollen when a single anther could be seen, and plants were recorded as silking, when one or more silks were visible. Anthesis-silking interval was calculated as the difference in days between male flowering and female flowering. Plant and ear height were recorded two weeks after pollination: plant height was the height (cm) from the soil surface to the flag leaf collar and ear height was the height (cm) from the soil surface to the stalk node at which the uppermost ear has emerged. The flag leaf angle was recorded using a protractor. The protractor was placed against the portion of stalk beneath the flag leaf. The protractor was held underneath the flag leaf's midrib to record the flag leaf angle at the point of attachment to the stalk. Tassel length was measured two weeks after pollination as the length (cm) between the flag leaf node and the top of the tassel. The number of primary tassel branches was recorded simultaneously as tassel length by counting the number of primary tassel branches that branch directly off the main branch.

2.4 Statistical data analysis

Data were analyzed with the following linear model:

$$Y_{ijklmn} = \mu + E_i + R(E)_{li} + G_j + GE_{ij} + D(G)_{jk} + ED(G)_{ijk} + P(ER)_{mil} + A(ER)_{nil} + \epsilon_{ijklmn}$$

where: Y_{ijklm} is the response in the environment *i*, group *j*, DH line *k*, replicate block *l*, pass *m* (i.e., field rows), range *n* (i.e., field columns); μ is the overall mean; E_i is the effect of environment *i*; $R(E)_{li}$ is the effect of replicate block *l* within environment *i*; G_j is the effect of the group of DH line *j*; GE_{ij} is the effect of the interaction between group *j* and environment *i*; $D(G)_{jk}$ is the effect of the DH line *k* within the group *j*; $ED(G)_{ijk}$ is the effect of the interaction between environment *i* and DH line *k* within the group of DH line *j*; $P(ER)_{mil}$ is the effect of the pass *m* within the environment *i* and replication *l*; $A(ER)_{nil}$ is the effect of the range *n* within the environment *i* and replication *l* and ϵ_{ijklm} is the effect of the residual error of the range *n*, pass *m*, block *l*, individual DH line *k*, group of DH line *j* and environment *i*. The effects of the environment, replicate block within environment, group of DH lines were considered fixed effects. All other effects were considered random. All phenotypic data analyses were conducted using the MIXED procedure of SAS 9.4 software (SAS Institute, Cary, NC). After fitting the full linear model to all traits, data were checked for outliers by computing the probability of studentized residuals using the t-distribution and adjusted with a Bonferroni correction for the number of residuals. Observations were considered outliers if the Bonferroni corrected P-value on the residuals were below 0.02. Then, a model containing all fixed effects but with different combinations of the random effects and homogeneity/heterogeneity in the residual variance across environments was tested for each trait.

Based on the smallest Bayesian Information Criteria (BIC; Schwarz, 1978), we decided which random effects to retain in the model. A final model was identified as having the best fit for each trait. The model with the smallest BIC value is shown in

supplemental materials (Supplementary Table 1). Variance components were estimated by REML (Patterson and Thompson, 1971), and likelihood ratio tests were performed to verify the significance of them. Overall means of the DH line groups were compared using Tukey's honest significant difference (HSD) procedure. Supplementary Table 2 shows BLUP values for 132, 185 and 170 DH lines from C0_DHL, C17_DHL and C0/C17_DHL groups, respectively. This information should be used to identify DH lines with both significant C0 background and modern plant architecture traits conferring adaptation to high plant density.

Repeatability was calculated with the formula:

$$\text{Repeatability} = \frac{\hat{\sigma}^2_{D(G)}}{\hat{\sigma}^2_{D(G)} + \frac{\hat{\sigma}^2_{ED(G)}}{e} + \frac{\hat{\sigma}^2_{\epsilon}}{re}}$$

where $\hat{\sigma}^2_{D(G)}$ corresponds to the variance estimate due to the DH line within group effect, $\hat{\sigma}^2_{ED(G)}$ is the variance estimate due to the interaction between environment and DH line, $\hat{\sigma}^2_{\epsilon}$ is the residual variance estimate and r and e are the number of replications and environments, respectively (Carena et al., 2010). The Pearson correlation coefficients between BLUPs from $D(G)_{jk}$ effect were calculated using the R software (R Core Team, 2021).

2.5 Genotyping and quality control

Genomic DNA was extracted from each DH line seedling established in the greenhouse at the Department of Agronomy, ISU. Leaf tissue samples from three plants per DH line were collected at the 3-4 leaf developmental stage, and the DNA extraction was done using the standard CIMMYT laboratory protocol (CIMMYT, 2005). Genotyping was carried out using the Diversity Arrays Technology sequencing (DArT-seq) method (Kilian et al., 2012) provided by the Genetic Analysis Service for Agriculture (SAGA) at CIMMYT. DArT-seq is a high-throughput, robust, reproducible, and cost-effective marker system based on genome complexity reduction using a combination of restriction enzymes, followed by hybridization to microarrays to simultaneously assay hundreds to thousands of markers across the genome (Sansaloni et al., 2011).

A total of 51,418 SNP markers were generated, but only 32,929 SNP markers were successfully called within the B73 RefGen_v5. The 32,929 SNP markers were filtered according to the following criteria: 1) Minimum call rate, 2) Minor Allele Frequency (MAF), 3) duplicated and monomorphic markers, and 4) heterozygosity. We used a threshold of ≥ 50% to remove poorly genotyped SNP markers, for which information was missing for more than half of the lines. SNP markers with MAF ≤ 1% were excluded. Duplicated and monomorphic SNP markers were removed using conditional formatting in Excel. Finally, genotypes with significant heterozygosity (not expected in DH or inbred lines) were excluded. After filtering and quality control, 13,846 SNP markers remained. In total, 29 DH lines (3 in C0_DHL, 9 in C17_DHL, and 17 in C0/C17_DHL) were discarded from the GWAS analysis due to obvious phenotypic segregation observed in field trials or missing genotypic or phenotypic data.

The software TASSEL v.5.2.70 (Bradbury et al., 2007) was used for the imputation of missing data using the LDkNNi (linkage disequilibrium k-nearest neighbors imputation) method (Money et al., 2016). LDkNNi process considers the linkage disequilibrium (LD) between SNPs when choosing the nearest neighbors. It exploits the fact that markers useful for imputation are often not physically close to the missing genotype rather distributed throughout the genome (Money et al., 2016).

2.6 Linkage disequilibrium and population structure

The average LD decay between SNP markers for each chromosome was determined in each group of DH lines using the squared Pearson correlation coefficient (r^2) between alleles at two loci for all possible combinations of alleles, and then weighting them according to the allele frequency. P-values were determined by a two-sided Fishers Exact test (Bradbury et al., 2007). The option "Full Matrix LD" on TASSEL v.5.2.70 was used to calculate LD for every combination of sites in the alignment (Bradbury et al., 2007). The resulting data were imported into R (R Core Team, 2021) to create LD decay plots and fit a smooth line using Hill and Weir expectations of r^2 between adjacent sites (Hill and Weir, 1988).

The selected 487 DH lines were known to belong to the three subpopulations BSSS, BSSS(R)C17, and BSSS/BSSS(R)C17. A principal component analysis (PCA) was conducted for all DH lines using the software GAPIT v.3 (Lipka et al., 2012). The principal components, plotted in a two-dimensional plot using discriminant analysis of principal components (DAPC), correctly identified a clear grouping of the DH lines into the three groups (C0_DHL, C17_DHL and C0/C17_DHL). The first two principal components explained 14.3% of the total SNP variation in the entire panel. Also, the C0/C17_DH lines group was scattered over a broader range, similar to the C0_DHL group. PCA results and molecular characterization of the DH lines within and among the cycles of selection are presented in Ledesma et al. (2023). The incorporation of population structure through PCA as a covariate in the fixed effect model increases the power to detect associations, and it has the advantage of eliminating false positives due to non-genetic effects associated with the population structure.

2.7 Genome-wide association studies

For GWAS analyses, we used four phenotypic traits that are known to be associated with adaptation to high plant density: male and female flowering, flag leaf angle, and the number of primary tassel branches. GWAS analysis was performed for each subpopulation individually (C0_DHL, C17_DHL, and C0/C17_DHL) and for the entire panel (487 DH lines) in order to determine how to best use DH lines for GWAS. The software package GAPIT (Lipka et al., 2012) was used for GWAS analysis. The fixed and random model circulating probability unification (FarmCPU) method was implemented in GAPIT. FarmCPU includes PCA results as a covariate, kinship as an additional

covariate to account for the relatedness among individuals (VanRaden, 2008), and additional algorithms that aid in solving the confounding problem between testing markers and covariates (Liu et al., 2016).

The P-values from each respective SNP were adjusted using False Discovery Rate (FDR) according to the Benjamini and Hochberg method (Benjamini and Hochberg, 1995). This statistic is also known as q-value and represents the estimated FDR if the associated P-value is used to declare significance. The default significant threshold value implemented in GAPIT was set at FDR < 0.05. We used the uniform Bonferroni-corrected threshold of $\alpha = 0.05$ for the significance level. Therefore, the suggested P-value was computed with α/n (n = 13,846, total markers used), and we obtained a P-value threshold of 3.61×10^{-6} for GWAS. Manhattan plots were used to visualize the significance of SNPs by chromosome location across the whole genome for each trait. Allele frequencies within population were estimated for each significant SNP by using the *popgen* function from snpReady R package (Granato et al., 2018).

2.8 Candidate gene mining

The available maize genome sequence (B73; RefGen_v5) was used as the reference genome for candidate gene identification. Genes were considered as candidates if a significantly associated SNP marker with phenotypic variance explained (PVE) higher than 5% was located within the range of LD decay observed for each chromosome (upstream and downstream). Candidate genes were identified using the Ensembl Biomart tool (Kinsella et al., 2011) and checked according to the SNP marker's physical position in the MaizeGDB molecular marker database (http://www.maizegdb.org; Portwood et al., 2019). Functional annotations of candidate genes were predicted in NCBI (http://www.ncbi.nlm.nih.gov/gene) and were also compared to previously published candidate genes.

3 Results

3.1 Phenotypic data analyses

Descriptive statistical analysis confirmed trait variability in the different groups of DH lines (Table 1). Phenotypic differences ($P \leq 0.05$) for all traits, except plant height, were found among groups of DH lines. DH lines within the C0_DHL group had the highest mean values for flowering time, ear height, flag leaf angle, tassel length and the number of primary tassel branches and were found to be different ($P \leq 0.05$) between the C17_DHL and C0/C17_DHL groups. On the other hand, DH lines within C17 group had the lowest values for these traits (Table 1). The C17_DHL group had the smallest anthesis-silking interval (0.1), meaning that plants showed silks and pollen shed almost simultaneously. Variance components due to DH lines within group effect were significant ($P< 0.05$) by the likelihood ratio test for all traits. Repeatabilities calculated for the complete set of DH lines across the three locations were found to be high across all traits. They ranged from 0.82 to 0.94 (Table 1). The correlation between the BLUPs were explored to determine relationships among evaluated traits (Table 2). The closest positive correlation (r = 0.88) was observed between male flowering and female flowering ($P \leq 0.001$). Plant and ear height were significantly ($P \leq 0.001$) and positively correlated (r = 0.76). They were also significantly and positively correlated with almost all other traits, except for the number of primary tassel branches and anthesis–silking interval.

3.2 Linkage disequilibrium

LD decay varied across the ten chromosomes and different regions within chromosomes (Figure 1). The C17_DHL group showed the largest LD decay distance ranging from 1,067 to 2,218 kb on chromosomes 5 and 4, respectively (Table 3). In contrast, the C0/C17_DHL group displayed the smallest LD decay distance (from 284 kb on chromosome 10 to 653 kb on chromosome 3). For C0_DHL, the LD decay ranged from 377 to 848 kb on chromosomes 10 and 3, respectively. The genome-wide LD decay distance was 569 kb, 1,509 kb and 463 kb for the C0_DHL, C17_DHL and C0/C17_DHL groups, respectively (Table 3). The genome-wide LD decay distance over all ten chromosomes in the entire panel of DH line panel was equal to 555 kb.

3.3 Genome-wide association studies

In total, 26 significant SNP markers were identified by FarmCPU (Table 4). A greater number of significant SNPs was found when the entire panel of DH lines (487 DH lines) was combined and used for GWAS with FarmCPU model. Therefore, the associations from the entire panel were considered for further analyses. A total of 22 SNP markers were found significant when using FarmCPU model with the entire panel of DH lines. Among those, two and one SNP presented PVE higher than 5% for flag leaf angle and number of primary tassel branches, respectively (Figure 2; Supplementary Table 3). No significant SNP was detected for male flowering trait (Table 4). By searching for candidate genes up and downstream for those three SNP markers being in LD with the corresponding chromosome based on the B73 RefGen_v5, 19 candidate genes were identified (Table 5). Ten candidate genes were identified for flag leaf angle and nine for number of primary tassel branches. We observed that for most significant SNPs, the allele frequencies were lower within C0_DHL, intermediate within C0C17_DHL and highest within C17_DHL population (Supplementary Table 4).

4 Discussion

4.1 Plant architecture traits adapting to high plant density

The breeding potential of the BSSS maize population DH lines is reflected by the distribution of the plant architecture traits that

10.3389/fpls.2023.1294507

TABLE 1 Statistics of flowering and plant architecture traits in different groups of DH lines derived from the BSSS maize population.

Trait	Group[a]	Mean	$\hat{\sigma}^2_{D(G)}$	Repeatability
Male flowering (days)	C0_DHL	67.4 a	6.8*	0.94
	C17_DHL	62.7 c	4.2*	
	C0/C17_DHL	65.5 b	5.2*	
Female flowering (days)	C0_DHL	69.1 a	8.5*	0.94
	C17_DHL	62.7 c	4.7*	
	C0/C17_DHL	66.4 b	7.7*	
Anthesis-silking interval (days)	C0_DHL	-1.7 a	1.6*	0.82
	C17_DHL	0.1 c	0.8*	
	C0/C17_DHL	-0.9 b	1.6*	
Plant height (cm)	C0_DHL	169.4 a	274.2*	0.93
	C17_DHL	170.5 a	194.4*	
	C0/C17_DHL	172.6 a	246.3*	
Ear height (cm)	C0_DHL	83.5 a	216.5*	0.92
	C17_DHL	68.5 c	122.4*	
	C0/C17_DHL	79.6 b	212.2*	
Flag leaf angle (Degrees from vertical)	C0_DHL	42.2 a	158.2*	0.89
	C17_DHL	13.8 c	48.1*	
	C0/C17_DHL	30.6 b	146.8*	
Tassel length (cm)	C0_DHL	42.1 a	16.1*	0.90
	C17_DHL	36.9 c	14.7*	
	C0/C17_DHL	38.9 b	21.1*	
Primary tassel branches (number)	C0_DHL	15.4 a	1.5*	0.94
	C17_DHL	7.3 c	0.4*	
	C0/C17_DHL	10.4 b	0.8*	

[a]Group, C0_DHL corresponds to the 132 derived DH lines from cycle 0, C0/C17_DHL corresponds to the 170 derived DH lines from C0/C17, and C17 corresponds to the 187 derived DH lines from cycle 17. Mean values were estimated from trait BLUPs of n lines within each group; $\hat{\sigma}^2_{D(G)}$ = variance estimate due to DH lines within group effect; * significant at 0.01 by the likelihood ratio test.

Means with the same letter in column are not statistically different at the 0.05 level of probability using Tukey's HSD comparison.

TABLE 2 Pearson correlation coefficients (r) between BLUPs for flowering and plant architecture traits of DH lines developed from the BSSS maize population.

	MAFL	FEFL	ASI	PLHE	EAHE	FLA	TALE	NPTB
MAFL	1							
FEFL	0.88**	1						
ASI	-0.02	-0.48	1					
PLHE	0.20**	0.14**	0.06	1				
EAHE	0.36**	0.27**	0.10*	0.76**	1			
FLA	-0.04	-0.05	0.02	0.10*	0.15*	1		
TALE	-0.05	0.01*	-0.11	0.24**	0.13*	-0.07	1	
NPTB	0.02	0.08*	-0.12	-0.02	0.07	0.10*	-0.04	1

** Significant at P ≤ 0.001, * Significant at P ≤ 0.05.

MAFL, male flowering; FEFL, female flowering; ASI, anthesis–silking interval; PLHE, plant height; EAHE, ear height; FLA, flag leaf angle; TALE, tassel length; NPTB, number of primary tassel branches.

10.3389/fpls.2023.1294507

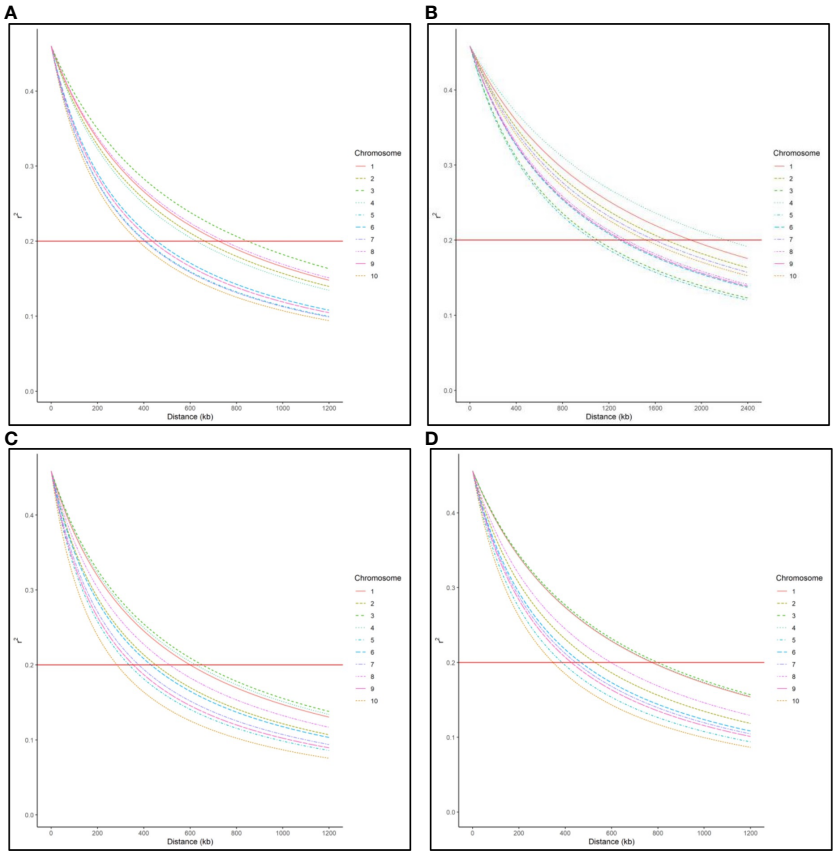


FIGURE 1
Genome-wide LD decay distance in **(A)** C0_DHL group, **(B)** C17_DHL group, **(C)** C0/C17_DHL group, and **(D)** Entire panel of 487 DH lines.

have been modified in this population, and these traits are involved in the adaptation to high plant densities (Duncan et al., 1967; Duncan, 1971; Mock and Pearce, 1975; Brekke et al., 2011). C17_DHL group presented the most favourable traits when adapting germplasm to higher plant densities (Table 1), such as reduced anthesis-silking interval, more erect leaves, and fewer primary tassel branches. The phenotypic data used in our study showed high values of repeatability, ranging from 0.82 to 0.94. These repeatabilities values agree with other studies (Buckler et al., 2009; Romay et al., 2013; Peiffer et al., 2014; Vanous et al., 2018).

In this study, we found significance differences in the mean of the plant architecture traits among the group of DH lines and a

TABLE 3 Linkage disequilibrium decay distance (kb) per chromosome in the different groups of DH lines and the entire panel.

Chromosome	C0_DHL	C17_DHL	C0/C17_DHL	Entire panel
1	724	1,911	600	774
2	663	1,698	452	529
3	848	1,104	653	799
4	627	2,218	625	781
5	408	1,067	336	386
6	456	1,298	431	467
7	404	1,597	379	445
8	748	1,352	512	597
9	436	1,320	355	426
10	377	1,523	284	348
Genome-wide LD	569	1,509	463	555

TABLE 4 The number of significant SNP markers associated with flowering and plant architecture traits in different groups of DH lines and the entire panel using FarmCPU model.

Population	Phenotypic traits				
	MAFL	FEFL	FLA	NPTB	Total
C0_DHL	0	0	0	0	0
C17_DHL	0	4	0	0	4
C0/C17_DHL	0	0	0	0	0
Entire panel	0	3	7	12	22
Total	0	7	7	12	26

MAFL, male flowering; FEFL, female flowering; FLA, flag leaf angle; NPTB, number of primary tassel branches.

reduction in the variance component estimates from the C17_DHL to the C0_DHL. Reduced genetic variance within the population was expected after 17 cycles of recurrent selection with recombination of a finite number of lines (10 or 20) within each cycle of selection. Flowering time showed a reduction of four days to anthesis and six days to silking from C0_DHL to C17_DHL groups. However, all DH lines flowered within a timeframe expected for the central US Corn Belt. Reduction in flag leaf angle has been reported in hybrids through the selection process and adaptation to high plant density (Duvick, 2005), as we found in this study. C17_DHL group could be a source of favourable alleles that impact more erect flag leaf angles. Additionally, we found a reduction in the number of primary tassel branches from an average of 15 in C0_DHL to 7 in the C17_DHL groups. These results confirmed a reduction in the number of primary tassel branches found by Edwards (2011) in the recurrent selection in the BSSS maize population. Additionally, these results are also in agreement with Brekke et al. (2011), where changes in plant architecture traits such as more upright flag leaf angle and reduction on the number of tassel branches were found as the cycles of selection advanced in the BSSS maize population. Large tassels can intercept enough light to lower photosynthetic rates in the canopy (Duncan et al., 1967), suggesting that smaller tassels may be advantageous for light utilization. However, Duncan et al. (1967) pointed out that this does not necessarily preclude some benefit of improved assimilate allocation with smaller tassels.

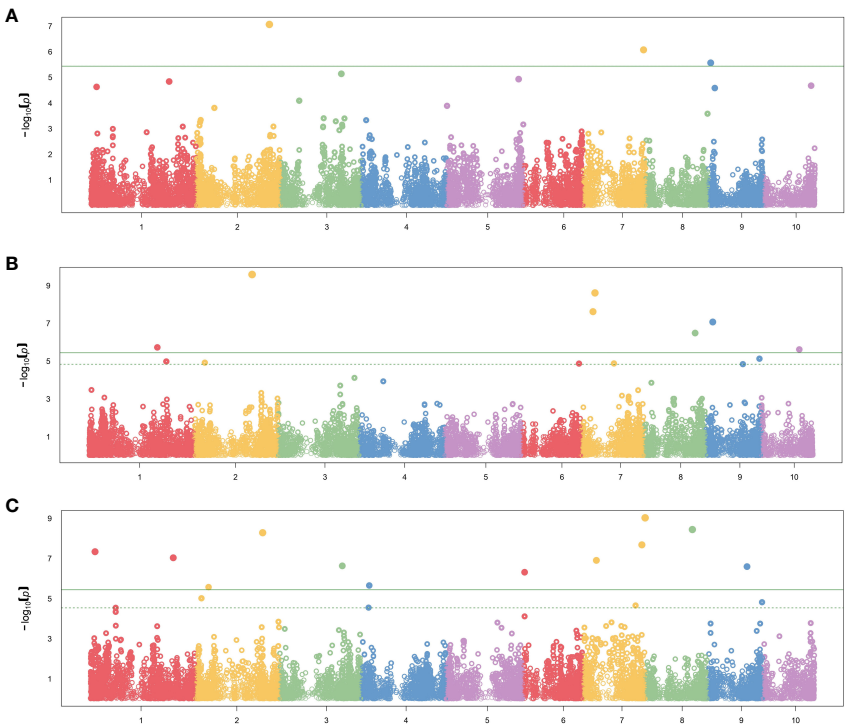


FIGURE 2
Manhattan plot results showing significant SNP markers associated with **(A)** female flowering, **(B)** flag leaf angle, **(C)** number of primary tassel branches in the entire panel using FarmCPU method. The X-axis plot represents the genomic position of the SNPs per chromosome. The Y-axis represents the negative logarithm of the P-value obtained from the GWAS model. The dash horizontal line represents the threshold from the FDR, and the solid horizontal line represents the threshold from the Bonferroni correction method.

10.3389/fpls.2023.1294507

TABLE 5 Candidate genes associated with plant architecture traits in the BSSS DH lines.

Traits	Chr	Gene start (Kbp)	Gene ID MaizeGDB	Gene ID Gramene	Gene name	Annotation
Flag leaf angle	1	199659	Zm00001eb036930	GRMZM2G136453	pap15	purple acid phosphatase15
	1	199726	Zm00001eb036940	GRMZM2G043198	pdh2	pyruvate dehydrogenase2
	1	199884	Zm00001eb036970	GRMZM2G159996	col16	C_2C_2-CO-like-transcription factor 16
	1	200045	Zm00001eb036990	GRMZM5G882527	bhlh173	bHLH-transcription factor 173
	1	200445	Zm00001eb037120	GRMZM2G033828	rrb3	retinoblastoma family3
	1	199275	Zm00001eb036880	GRMZM5G872141	sweet11	sugars will eventually be exported transporter11
	1	199317	Zm00001eb036890	GRMZM2G068657	pat4	protein S-acyltransferase4
	1	199461	Zm00001eb036910	GRMZM2G064962	gpx3	glycerophosphodiester phosphodiesterase3
	2	167752	Zm00001eb095620	GRMZM2G161382	cyc11	cyclin11
	2	168274	Zm00001eb095690	GRMZM2G087955	myb139	MYB-transcription factor 139
Number of primary tassel branches	2	194428	Zm00001eb101630	GRMZM2G022162	ca5p12	CCAAT-HAP5-transcription factor 512
	2	194465	Zm00001eb101670	GRMZM2G003992	mlkp3	Maize LINC KASH AtWIP-like3
	2	194543	Zm00001eb101700	GRMZM2G052671	wrky71	WRKY-transcription factor 71
	2	194575	Zm00001eb101720	GRMZM2G350857	abi53	ABI3-VP1-transcription factor 53
	2	194727	Zm00001eb101780	GRMZM2G080439	upl1	ubiquitin-protein ligase1
	2	194795	Zm00001eb101840	GRMZM2G134334	znf13	zinc finger protein13
	2	194921	Zm00001eb101880	GRMZM2G417597	bhlh6	bHLH-transcription factor 6
	2	195046	Zm00001eb101910	GRMZM2G125934	bzip85	bZIP-transcription factor 85
	2	195064	Zm00001eb101920	GRMZM2G126018	sbp23	SBP-transcription factor 23

Ear height for the C17_DHL group was significantly lower than for C0_DHL, which might be partially due to the inbreeding depression. Plant and ear height are traits of interest when adapting germplasm as they are closely associated with flowering time, lodging resistance, biomass production, and grain yield (Durand et al., 2012; Teng et al., 2013). By reducing height traits during the selection for industrial agriculture, it was observed an increased harvest uniformity, favourably partition carbon and nutrients between grain and non-grain biomass, and enhanced fertilizer, pesticide, and water use efficiency (Khush, 2001). C17_DHL group was altered in important traits for high plant density tolerance compared to the C0_DHL group. In general, the C17_DHL group showed a better performance in plant architecture traits than the C0_DHL group. These differences demonstrate that 17 cycles of recurrent selection have been effective. At the same time, the C0/C17 DHL group showed considerable variation and could be used as a source to develop DH lines and hybrids adapted to high planting densities. Developing DH lines in more advanced cycles of selection improved agronomic traits, such as flowering time, flag leaf angle, and the number of primary tassel branches. If there were few major loci available in early selection cycles, they probably got fixed during the selection process. Therefore, the extraction of DH lines out of the BSSS maize population was effective, as indicated by plant architecture traits that suggested adaptation to high plant density. Some correlations coefficients were significant, indicating that adaptation based on plant architecture traits is a viable option in altering other important adaptation-related traits.

4.2 The exploitation of early cycle of BSSS DH lines

A method to exploit maize's genetic diversity is introducing exotic germplasm and/or using landraces as a source of new alleles. However, several cycles of inbreeding are required. Additionally, inbreeding from landraces results in a high load of recessive alleles, mutations, and deleterious alleles that need to be selected against by conventional breeding methods (Strigens et al., 2013). According to Ledesma et al. (2023), the 17 cycles of reciprocal recurrent selection program have left behind useful genetic variation present in the C0_DHL during the selection process. Thus, to have a sufficient number of lines to be evaluated for testcross performance from exotic germplasm or landraces, it is necessary to start the breeding program with a large number of plants. This laborious effort is the main reason why exotic germplasm and landraces are limited used in modern breeding programs (Goodman, 2005). However, DH technology can enable more effective access to the genetic diversity of landraces and exotic germplasm in a faster way (Strigens et al., 2013; Chaikam et al., 2019). In this context, the C0_DHL group

could be a reservoir of genetic diversity that could be untapped using DH technology. Deleterious alleles are expressed in the haploid stage and can be purged through selection. Hence, DH technology is a useful tool to access the genetic diversity present in landraces and to expand the genetic diversity of the elite germplasm (Wilde et al., 2010; Strigens et al., 2013; Böhm et al., 2017; Chaikam et al., 2019).

Developing DH lines from earlier cycles of recurrent selection programs could be an alternative approach to conventional breeding for introduction of diversity into related elite lines. In this study, we developed DH lines from the earlier cycle of the BSSS maize population to explore the phenotypic variation that has been left behind when advancing cycles of recurrent selection. Significant phenotypic variation was observed between the groups of DH lines for all traits evaluated, except for plant height. C17_DHL group presented the most favorable characteristics when adapting germplasm to higher plant densities (i.e., lowest means for flowering time, ear height, flag leaf angle, tassel length and the number of primary tassel branches). However, the genetic variability among the C0_DHL and the C0/C17_DHL allowed the identification of DH lines with desirable plant architecture traits that confer adaptation to high plant density. Some of these DH lines are a promising source of favorable alleles for plant density response. Thus, selected DH lines could be introgressed into current germplasm to improve the adaptation to high plant density. The large genetic distances of the C0_DHL compared to the C17_DHL (Ledesma et al., 2023) demonstrated the potential of the C0_DHL group to broaden the genetic base of the Stiff Stalk (SS) germplasm. However, more studies need to be conducted at the testcross level to know the hybrid combinations' performance. The use of early selected cycles and DH technology opens new opportunities for exploring genetic diversity in available germplasm.

4.3 Linkage disequilibrium and GWAS analysis

LD refers to the nonrandom association of alleles at different loci in a breeding population (Flint-Garcia et al., 2003). It can be estimated using the correlation between SNP markers. The magnitude of LD and its decay with the genetic distance is important to determine the resolution of association mapping because LD's extent determines the required number of SNP markers and the mapping resolution (Vos et al., 2017). In our entire panel of BSSS DH lines, we found that the LD decayed over 555 kb across the genome at the $r^2 = 0.2$ threshold (Figure 1D). However, LD decay varied across the ten chromosomes and different genetic regions within chromosomes ranging from 348 kb in chromosome 10 to 799 kb in chromosome 3 (Figure 1D). These results agree with Vanous et al. (2018). They investigated a diverse panel consisting of exotic derived DH lines and found that LD decayed over a distance greater than 500 kb for all chromosomes. The LD within the C17_DHL group is quite more extensive than in C0_DHL and C0/C17_DHL. The larger LD decay distance observed in the C17_DHL group may be due to the breeding history of the population (e.g., the occurrence of bottlenecks) and the lower genetic diversity represented by this population. LD decay is more rapid in pools with higher genetic diversity (Romay et al., 2013; Wu et al., 2016). The C17_DH lines came from a population that was gone through 17 cycles of recurrent selection, which have caused some genetic drift, or a small effective population size, resulting in the larger decay distances.

The rapid LD decay, together with high genotypic variances and absence of population structure within populations, enables good resolution association mapping in some germplasm (Strigens et al., 2013). In our study, when we analyzed each group of DH lines (C0_DHL, C17_DHL, and C0/C17 DHL) the number of SNP markers associated was low or absent. However, when we used the entire panel of DH lines, we found 22 SNP markers among all traits. These results could be due to the lower variation within each DH line group or the smaller population size that affects the power to detect associations. Another possible reason for having low power to identify associated SNP markers to plant architecture traits when we performed the analysis by each group of DH lines could be due to the fixation of alleles. In the C17_DHL group, there are major genes affecting plant architecture traits and respective alleles are present at a low frequency in the C0_DHL group.

Alleles were in higher frequency within the population C17_DHL (Supplementary Table 4). The intermediate allele frequencies observed within population C0C17_DHL suggests that this population might be the most powerful population for GWAS studies, as those lines segregate for favorable alleles. Most significant SNPs were detected when using the entire panel not only because of its population structure, but mainly because the sample size was higher when using the entire panel. It has been largely discussed that sample size plays an important role in GWAS studies (Ibrahim et al., 2020; Wang et al., 2020; Murphy et al., 2022). Therefore, we believe that an increased sample size of C0C17_DHL could increase its power of SNP detection.

4.4 Candidate genes for plant architecture traits adapting to high plant density

Since we found consistent changes in at least four traits that are known to be associated with adaptation to high plant density we focused our discussion on candidate genes for these traits. Trends for reducing male and female flowering time, the number of primary tassel branches, and more upright flag leaf angles in C17_DHL compared with the C0_DHL were identified in our work. These trends have been reported for parental inbred lines of hybrids previously released (Duvick, 2005; Lauer et al., 2012),

which could reflect a correlated response of modern breeding germplasm to selection for grain yield under higher plant densities (Edwards, 2011).

Flag leaf angle and number of primary tassel branches presented significant SNP markers with PVE higher than 5% (Supplementary Table 3). Thus, we searched potential candidate genes for these two traits. In total, 19 candidate genes were found. Flag leaf angle has experienced changes when advancing cycles in the recurrent selection program. In this study, we found that C17_DHL had a more upright flag leaf angle than the other two groups of DH lines. These results agree with different hybrids studies where a trend toward vertical flag leaf angle had been observed in recent decades. More vertical upper leaves are the desired trait since permit lighter to penetrate the canopy, improving the photosynthetic efficiency and allowing farmers to plant maize at higher densities (Edwards, 2011). In this study, an important region on chromosome 7 with PVE equal to 10.81% was identified. This suggests that the surrounding genomic region might have a strong association on modifying flag leaf angle, which could help to dramatically alter the trait.

Early studies conducted in maize to dissect the genetic basis of leaf angle have identified several quantitative trait loci and genomic regions for leaf angle throughout all the ten maize chromosomes. Dzievit et al. (2019) found 12 QTL on chromosome 1, 2, 3, 4 and 8 affecting leaf angle. Additionally, several genes have been cloned as the outcome of the combined use of quantitative genetics and induced or natural mutants associated with changes in leaf angle in maize (Mantilla-Perez and Fernandez, 2017). The number of primary tassel branches is considered as the principal component of maize tassel inflorescence architecture and is a typical quantitative trait controlled by multiple genes (Chen et al., 2017). Reductions in tassel size and tassel branch number have continuously decreased over time (Duvick, 2005). Previous studies in the BSSS maize population have revealed changes through advancing cycles in the recurrent selection program (Brekke et al., 2011). According to Duncan et al. (1967), tassels could block enough sunlight to reduce photosynthesis by 19%. We identified nine candidate genes controlling the number of primary tassel branches which will be useful for its improvement by molecular breeding and provide a basis for the cloning of the genes. Chen et al. (2017) identified 11 QTL located in chromosomes 2, 3, 5, and 7 demonstrating that tassel branch number variation was mainly caused by alleles with a major effect, minor effect, and slightly modified by epistatic effects.

DH lines developed in this study could be sources of new germplasm for broadening the genetic variation compared to elite germplasm to develop varieties or hybrids adapted to the US corn belt. Thus, individual lines with superior performance for agronomic and morphological traits can be selected and introgressed into elite materials. However, the testcross performance of the DH lines remains to be evaluated to test their yield potential in hybrid combinations. Additionally, in this study, we found that the entire panel of DH lines could be used for association analysis for flowering and plant architecture traits. Instead of using each DH line group individually, the power of detecting associated SNP increased when we used the entire panel of DH lines. Additionally, identifying QTL or regions for plant architecture traits in this study may help to elucidate the genetic basis of these traits and facilitate future work about marker-assisted selection or map-based cloning in maize breeding programs.

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: https://doi.org/10.25380/iastate.22893878.v1, Iowa State University DataShare, accession 22893878.

Author contributions

AL: Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. AS: Data curation, Formal analysis, Software, Validation, Writing – review & editing. FR: Writing – review & editing. FA: Data curation, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. JE: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing, Investigation. UF: Data curation, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. TL: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, 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 work was supported by USDA's National Institute of Food and Agriculture (grant numbers: IOW04714, IOW05520; IOW05510; IOW05656; NIFA award 2018-51181-28419 and 2020-51300-32180), US. Department of Agriculture, Agricultural Research Service, the Iowa State University Plant Sciences Institute, Iowa State University Crop Bioengineering Center, R.F. Baker Center for Plant Breeding, and the K.J. Frey Chair in Agronomy at Iowa State University.

Acknowledgments

This research was supported in part by the US. Department of Agriculture, Agricultural Research Service (USDA ARS) Project No. 5030-21000-073-000D. Mention of trade names or commercial

products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and Employer. Alejandro Ledesma Miramontes acknowledges the National Council for Science and Technology (CONACYT) and the National Institute for Agricultural, Livestock, and Forestry Research (INIFAP) in Mexico for the scholarship 2016 for Ph.D. studies.

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.2023.1294507/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 (1), e20014. doi: 10.1002/tpg2.20014

Andorf, C., Beavis, W. D., Hufford, M., Smith, S., Suza, W. P., Wang, K., et al. (2019). Technological advances in maize breeding: past, present and future. *Theor. Appl. Genet.* 132, 817–849. doi: 10.1007/s00122-019-03306-3

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to mutliple testing. *J. R. Stat. Soc B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Böhm, J., Schipprack, W., Utz, H. F., and Melchinger, A. E. (2017). Tapping the genetic diversity of landraces in allogamous crops with doubled haploid lines: a case study from European flint maize. *Theor. Appl. Genet.* 130 (5), 861–873. doi: 10.1007/s00122-017-2856-x

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

Brekke, B., Edwards, J., and Knapp, A. (2011). Selection and adaptation to high plant density in the Iowa Stiff Stalk Synthetic Maize (L.) population. *Crop Sc* 51 (5), 1965. doi: 10.2135/cropsci2010.09.0563

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., et al. (2009). The genetic architecture of maize flowering time. *Science* 325 (5941), 714–718. doi: 10.1126/science.1174276

Carena, M. J., Hallauer, A. R., and Miranda Filho, J. B. (2010). *Quantitative genetics in maize breeding* (New York, NY: Springer New York). doi: 10.1007/978-1-4419-0766-0

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 (12), 3227–3243. doi: 10.1007/s00122-019-03433-x

Chen, Z. J., Yang, C., Tang, D. G., Zhang, L., Zhang, L., Qu, J. T., et al. (2017). Dissection of the genetic architecture for tassel branch number by QTL analysis in two related populations in maize. *J. Integr. Agric.* 16 (7), 1432–1442. doi: 10.1016/S2095-3119(16)61538-1

CIMMYT (2005). *Laboratory protocols: CIMMYT applied molecular genetics laboratory. 3rd ed* (Mexico, D.F: CIMMYT).

Coffman, S. M., Hufford, M. B., Andorf, C. M., and Lübberstedt, T. (2019). Haplotype structure in commercial maize breeding programs in relation to key founder lines. *Theor. Appl. Genet.* 133, 547–561. doi: 10.1007/s00122-019-03486-y

Duncan, W. G. (1971). Leaf angles, leaf area, and canopy photosynthesis. *Crop Sc.* 11 (4), 482–485. doi: 10.2135/cropsci1971.0011183X001100040006x

Duncan, W. G., Loomis, R. S., Williams, W. A., and Hanau, R. (1967). A model for simulating photosynthesis in plant communities. *Hilgardia* 38 (4), 181–205. doi: 10.3733/hilg.v38n04p181

Durand, E., Bouchet, S., Bertin, P., Ressayre, A., Jamin, P., Charcosset, A., et al. (2012). Flowering time in maize: Linkage and epistasis at a major effect locus. *Genetics* 190 (4), 1547–1562. doi: 10.1534/genetics.111.136903

Duvick, D. N. (2005). The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv. Agron.* 86, 83–145. doi: 10.1016/S0065-2113(05)86002-X

Dzievit, M. J., Li, X., and Yu, J. (2019). Dissection of leaf angle variation in maize through genetic mapping and meta-analysis. *Plant Genome* 12 (1), 1–12. doi: 10.3835/plantgenome2018.05.0024

Eberhart, S. A., Debela, S., and Hallauer, A. R. (1973). Reciprocal recurrent selection in BSSS and BSCB1 maize populations and half-sib selection in BSSS. *Crop Sci.* 13, 451–456. doi: 10.2135/cropsci1973.0011183X001300040017x

Edwards, J. (2011). Changes in plant morphology in response to recurrent selection in the Iowa Stiff Stalk Synthetic maize population. *Crop Sc* 51 (6), 2352–2361. doi: 10.2135/cropsci2010.09.0564

Edwards, J. (2010). Testcross response to four cycles of half-sib and S-2 recurrent selection in the BS13 maize (*Zea mays L.*) Population. *Crop Sci.* 50, 1840–1847. doi: 10.2135/cropsci2009.09.0557

Flint-Garcia, S. A., Thornsberry, J. M., and Edward, S. B. IV (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Fritsche-Neto, R., Sabadin, F., DoVale, J. C., Borges, K. L. R., de Souza, P. H., Crossa, J., et al. (2023). Realized genetic gains *via* recurrent selection in a tropical maize haploid inducer population and optimizing simultaneous selection for the next cycles. *Crop Sci* 63, 2865–2876. doi: 10.1002/csc2.21081

Goodman, M. M. (2005). Broadening the U.S. maize germplasm base. *Maydica* 50 (3), 203–214.

Granato, I. S. C., Galli, G., Couto, E. G. de O., Souza, M. B., Mendonça, L. F., Fritsche-Neto, R., et al. (2018). snpReady: a tool to assist breeders in genomic analysis. *Mol. Breed.* 38, 102. doi: 10.1007/s11032-018-0844-8

Hallauer, A. R., and Darrah, L. L. (1985). Compendium of recurrent selection methods and their application. *Crit. Rev. Plant Sci.* 3 (1), 1–33. doi: 10.1080/07352688509382202

Helms, T. C., Hallauer, A. R., and Smith, O. S. (1989). Genetic drift and selection evaluated from recurrent selection programs in maize. *Crop Sci.* 29, 602–607. doi: 10.2135/cropsci1989.0011183X002900030009x

Hill, W. G., and Weir, B. S. (1988). Variances and covariances of squared linkage disequilibria in finite populations. *Theor. Popul. Biol.* 33 (1), 54–78. doi: 10.1016/0040-5809(88)90004-4

Hull, F. H. (1945). Recurrent selection for specific combining ability in corn. *Agron. J.* 37 (2), 134–145. doi: 10.2134/agronj1945.00021962003700020006x

Ibrahim, A. K., Zhang, L., Niyitanga, S., Afzal, M. Z., Xu, Y., Zhang, L., et al. (2020). Principles and approaches of association mapping in plant breeding. *Trop. Plant Biol.* 13, 212–224. doi: 10.1007/s12042-020-09261-4

Jenkins, M. T. (1940). The segregation of genes affecting yield of grain in maize. *Agron. J.* 32 (1), 55–63. doi: 10.2134/agronj1940.00021962003200010008x

Keeratinijakal, V., and Lamkey, K. R. (1993). Responses to reciprocal recurrent selection in BSSS and BSCB1 maize populations. *Crop Sci.* 33, 73–77. doi: 10.2135/cropsci1993.0011183X003300010012x

Khush, G. S. (2001). Green revolution: The way forward. *Nat. Rev. Genet.* 2 (10), 815–822. doi: 10.1038/35093585

Kilian, A., Wenzl, P., Huttner, E., Carling, J., Xia, L., Blois, H., et al. (2012). "Diversity arrays technology: A generic genome profiling technology on open platforms," in *Data production and analysis in population genomics. Methods in molecular biology*, vol. 888 . Eds. F. Pompanon and A. Bonin (Totowa, NJ: Humana Press). doi: 10.1007/978-1-61779-870-2_5

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011, bar030. doi: 10.1093/database/bar030

Lamkey, K. R. (1992). 50 years of recurrent selection in the Iowa Stiff Stalk Synthetic maize population. *Maydica* 37, 19–28.

Lauer, S., Hall, B. D., MuLaosmanovic, E., Anderson, S. R., Nelson, B., and Smith, S. (2012). Morphological changes in parental lines of Pioneer brand maize hybrids in the U.S. central Corn Belt. *Crop Sc.* 52 (3), 1033–1043. doi: 10.2135/cropsci2011.05.0274

Ledesma, A., Ribeiro, F. A. S., Uberti, A., Edwards, J., Hearne, S., Frei, U., et al. (2023). Molecular characterization of doubled haploid lines derived from different cycles of the Iowa Stiff Stalk Synthetic (BSSS) maize population. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1226072

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, 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 (2), e1005767. doi: 10.1371/journal.pgen.1005767

Mantilla-Perez, M. B., and Fernandez, M. G. S. (2017). Differential manipulation of leaf angle throughout the canopy: Current status and prospects. *J. Exp. Bot.* 68, 5699–5717. doi: 10.1093/jxb/erx378

Martin, J. M., and Hallauer, A. R. (1980). Seven cycles of reciprocal recurrent selection in BSSS and BSCB1 maize populations. *Crop Sci.* 20, 599–603. doi: 10.2135/cropsci1980.0011183X002000050013x

Mock, J. J., and Pearce, R. B. (1975). An ideotype of maize. *Euphytica* 24 (3), 613–623. doi: 10.1007/BF00132898

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G., and Myles, S. (2016). LinkImpute: fast and accurate genotype imputation for non-model. *G3 Genes|Genomes|Genetics* 5 (11), 2383–2390. doi: 10.1534/g3.115.021667

Murphy, M. D., Fernandes, S. B., Morota, G., and Lipka, A. E. (2022). Assessment of two statistical approaches for variance genome-wide association studies in plants. *Heredity* 129 (2), 93–102. doi: 10.1038/s41437-022-00541-1

Ordas, B., Butron, A., Alvarez, A., Revilla, P., and Malvar, R. A. (2012). Comparison of two methods of reciprocal recurrent selection in maize (Zea mays L.). *Theor. Appl. Genet.* 124, 1183–1191. doi: 10.1007/s00122-011-1778-2

Patterson, H. D., and Thompson, R. (1971). Recovery of inter-block information when block size are unequal. *Biometrics* 58, 545–554. doi: 10.1093/biomet/58.3.545

Peiffer, J. A., Romay, M. C., Gore, M. A., Flint-Garcia, S. A., Zhang, Z., Millard, M. J., et al. (2014). The genetic architecture of maize height. *Genetics* 196 (4), 1337–1356. doi: 10.1534/genetics.113.159152

Penny, L. H., and Eberhart, S. A. (1971). 20 years of reciprocal recurrent selection with 2 synthetic varieties of maize (*Zea mays* L.). *Crop Sci.* 11, 900–903. doi: 10.2135/cropsci1971.0011183X001100060041x

Pollak, L. M. (2003). The History and Success of the public–private project on germplasm enhancement of maize (GEM). *Adv. Agron.* 78, 46–89. doi: 10.1016/S0065-2113(02)78002-4

Portwood, J. L., Woodhouse, M. R., Cannon, E. K., Gardiner, J. M., Harper, L. C., Schaeffer, M. L., et al. (2019). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.* 47 (D1), D1146–D1154. doi: 10.1093/nar/gky1046

R Core Team (2021). *R: A language and environment for statistical computing* (R Foundation for Statistical Computing).

Romay, M. C., Flint-Garcia, S. A., Casstevens, T. M., Glaubitz, J. C., McMullen, M. D., Holland, J. B., et al. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol.* 14, R55. doi: 10.1186/gb-2013-14-6-r55

Sansaloni, C., Petroli, C., Jaccoud, D., Carling, J., Detering, F., Grattapaglia, D., et al. (2011). Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proc.* 5 (Suppl 7), 54. doi: 10.1186/1753-6561-5-s7-p54

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6 (2), 461–464. doi: 10.1214/aos/1176344136

Smith, O. S. (1983). Evaluation of recurrent selection in BSSS, BSCB1, and BS13 maize populations. *Crop Sci.* 23, 35–40. doi: 10.2135/cropsci1983.0011183X002300010011x

Sprague, G. F. (1946). Early testing of inbred lines of corn. *J. Am. Soc Agron.* 38 (2), 108–117. doi: 10.2134/agronj1946.00021962003800020002x

Strigens, A., Schipprack, W., Reif, J. C., and Melchinger, A. E. (2013). Unlocking the genetic diversity of maize landraces with doubled haploids opens new avenues for breeding. *PloS One* 8 (2), 7–9. doi: 10.1371/journal.pone.0057234

Teng, F., Zhai, L., Liu, R., Bai, W., Wang, L., Huo, D., et al. (2013). ZmGA3ox2, a candidate gene for a major QTL, qPH3.1, for plant height in maize. *Plant J.* 73 (3), 405–416. doi: 10.1111/tpj.12038

Vanous, A., Gardner, C., Blanco, M., Martin-Schwarze, A., Lipka, A. E., Flint-Garcia, S., et al. (2018). Association mapping of flowering and height traits in germplasm enhancement of maize doubled haploid (GEM-DH) lines. *Plant Genome* 11 (2), 170083. doi: 10.3835/plantgenome2017.09.0083

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.* 11, 147–157. doi: 10.1002/cppb.20050

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91 (11), 4414–4423. doi: 10.3168/jds.2007-0980

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., and van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* 130 (1), 123–135. doi: 10.1007/s00122-016-2798-8

Wang, Q., Tang, J., Han, B., and Huang, X. (2020). Advances in genome-wide association studies of complex traits in rice. *Theor. Appl. Genet.* 133, 1415–1425. doi: 10.1007/s00122-019-03473-3

Wilde, K., Burger, H., Prigge, V., Presterl, T., Schmidt, W., Ouzunova, M., et al. (2010). Testcross performance of doubled-haploid lines developed from European flint maize landraces. *Plant Breed* 129 (2), 181–185. doi: 10.1111/j.1439-0523.2009.01677.x

Wu, Y., San Vicente, F., Huang, K., Dhliwayo, T., Costich, D. E., Semagn, K., et al. (2016). Molecular characterization of CIMMYT maize inbred lines with genotyping-by-sequencing SNPs. *Theor. Appl. Genet.* 129 (4), 753–765. doi: 10.1007/s00122-016-2664-8

Zhao, Y., Wang, H., Bo, C., Dai, W., Zhang, X., Cai, R., et al. (2019). Genome-wide association study of maize plant architecture using F1 populations. *Plant Mol. Biol.* 99, 1–15. doi: 10.1007/s11103-018-0797-7

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 21 December 2023
DOI 10.3389/fpls.2023.1298591

OPEN ACCESS

EDITED BY
Mulatu Geleta,
Swedish University of Agricultural Sciences, Sweden

REVIEWED BY
Barbara Pipan,
Agricultural institute of Slovenia, Slovenia
Mian Abdur Rehman Arif,
Nuclear Institute for Agriculture and Biology, Pakistan

*CORRESPONDENCE
Mónica Mathias-Ramwell
monica.mathias@inia.cl

RECEIVED 21 September 2023
ACCEPTED 05 December 2023
PUBLISHED 21 December 2023

CITATION
Mathias-Ramwell M, Pavez V, Meneses M, Fernández F, Valdés A, Lobos I, Silva M, Saldaña R and Hinrichsen P (2023) Phenotypic and genetic characterization of an *Avena sativa* L. germplasm collection of diverse origin: implications for food-oat breeding in Chile. *Front. Plant Sci.* 14:1298591. doi: 10.3389/fpls.2023.1298591

Phenotypic and genetic characterization of an *Avena sativa* L. germplasm collection of diverse origin: implications for food-oat breeding in Chile

Mónica Mathias-Ramwell[1*], Valentina Pavez[2], Marco Meneses[2], Feledino Fernández[1], Adriana Valdés[3], Iris Lobos[4], Mariela Silva[4], Rodolfo Saldaña[5] and Patricio Hinrichsen[2]

[1]Programa de mejoramiento genético de avena, Instituto de Investigaciones Agropecuarias (INIA), Centro Regional de Investigación Carillanca, Temuco, Chile, [2]Laboratorio de Análisis Genético, Instituto de Investigaciones Agropecuarias, Centro Regional de Investigación La Platina, Santiago, Chile, [3]Facultad de Recursos Naturales, Universidad Católica de Temuco, Temuco, Chile, [4]Laboratorio de Espectroscopía Infrarrojo Cercano, Instituto de Investigaciones Agropecuarias, Centro Regional de Investigación Remehue, Osorno, Chile, [5]Laboratorio de Nutrición Animal y Medio Ambiente, Instituto de Investigaciones Agropecuarias, Centro Regional de Investigación Remehue, Osorno, Chile

Oats are known for their nutritional value and also for their beneficial properties on human health, such as the reduction of cholesterol levels and risk of coronary heart disease; they are an important export product for Chile. During the last decade (2010-2022) over 90% of the oat cultivated area in Chile has been covered with *Avena sativa* L. cv. Supernova INIA. This lack of genetic diversity in a context of climate change could limit the long-term possibility of growing oats in Chile. The present study is a phenotypic and genetic analysis of 132 oat cultivars and pure lines of diverse origin that can be considered as potential breeding material. The germplasm was evaluated for 28 traits and analyzed with 14 SSR markers. The effects of genotypes on phenotype were significant over all traits ($P \leq 0.05$). Most traits exhibited moderate to high broad-sense heritability with exceptions such as yield (H^2 = 0.27) and hulls staining (H^2 = 0.32). Significant undesirable correlations between traits were generally of small biological importance, which is auspicious for achieving breeding objectives. Some of the heritability data and correlations provided here have not been previously reported. The overall phenotypic diversity was high (H' = 0.68 $\pm$ 0.18). The germplasm was grouped into three phenotypic clusters, differing in their qualities for breeding. Twenty-six genotypes outperforming Supernova INIA were identified for breeding of conventional food-oats. The genetic diversity of the germplasm was moderate on average (He = 0.58 $\pm$ 0.03), varying between 0.32 (AM22) and 0.77 (AME178). Two genetic subpopulations supported by the Structure algorithm exhibited a genetic distance of 0.24, showing low divergence of the germplasm. The diversity and phenotypic values found in this collection of oat genotypes are promising with respect to obtaining genetic gain in the short term in breeding programs. However, the similar genetic diversity, higher phenotypic diversity, and better phenotypic

performance of the germplasm created in Chile compared to foreign germplasm suggest that germplasm harboring new genetic diversity will be key to favor yield and quality in new oat cultivars in the long term.

KEYWORDS

oat breeding, genetic diversity, phenotypic diversity, heritability, SSRs

1 Introduction

Oats are increasingly gaining popularity due to a rise in awareness among consumers about their nutritional value and health benefits; a compound annual growth rate of 4.38% in consumption in the 2023-2028 period is projected in Latin America (Informe de expertos, 2023). The consumption of oat-based products in specific quantities has positive effects in reducing cholesterol levels, risk of cancer and coronary heart disease, as part of a diet low in cholesterol and saturated fat (FDA, 2023). Oats occupy second place in total area of crops grown in Chile, with around 100,000 ha, and a national average yield ranging between 4.5 and 5 tons per hectare, with production mainly destined for export (ODEPA, 2023).

Chilean oat production in the last decade has been based mostly on one genotype, cv. Supernova INIA, a cultivar created by the New Zealand Institute for Plant and Food Research Limited and registered in Chile in 2010. Although a few locally developed cultivars have been released during the last decades, they have not affected the varietal turnover, since 70 to 90% of the oat area is still occupied by Supernova INIA (De la Fuente, 2022), resulting in crop genetic uniformity. The loss of crop genetic diversity in a given area over a period of time, measured through the decline in cultivar number, represents crop genetic erosion that can result in vulnerability, creating the potential for widespread crop failure (Khoury et al., 2022). This is crucial considering that adverse conditions for cultivation of crops, including biotic and abiotic stresses, are expected to increase with climate change and global warming (Ristaino et al., 2021; Skendžić et al., 2021; FAO, 2022).

The focus of the INIA Chile oats breeding program has been the creation of advanced lines with high yield and stability in diverse environments, adequate industrial grain quality for processing, tolerance to lodging and diseases, among others (Beratto, 2006; Mathias-Ramwell et al., 2016). Knowledge of the genetic and phenotypic diversity of the available germplasm would allow a proper conservation of the germplasm for the future and use in the creation of new cultivars. Broadening the genetic background of a species, improving or at least preserving the most relevant economic traits, is key for the development of environmentally resilient new cultivars (Swarup et al., 2021). Modern oat cultivars have exhibited a narrower gene pool than landraces (Montilla-Bascón et al., 2013; Cieplak et al., 2021). Genetic diversity also varies depending on geographic origin, for example oat germplasm from North and South America showed higher genetic diversity than those from Europe, as European genotypes are closely related to each other (Achleitner et al., 2008).

Understanding the genetic diversity of the breeding germplasm, in combination with the phenotypic variation, correlation and heritability of important traits under selection, has facilitated the design of more effective breeding programs (Nava et al., 2010; Krishna et al., 2013; Boczkowska et al., 2016; Winkler et al., 2016; Silveira et al., 2020; Cieplak et al., 2021). There is published research on this aspect in Sweden, Denmark, Finland, Norway (Nersting et al., 2006), USA (Winkler et al., 2016), India (Chauhan and Singh, 2019), and Brazil (Mazurkievicz et al., 2019; Meira et al., 2019; Zimmer et al., 2019; Silveira et al., 2020), but not under the environmental conditions of southern Chile. Few Chilean oat genotypes have been included in genetic diversity studies (Achleitner et al., 2008), which do not include the newest breeding lines and those most cultivated in Chile.

Different types of molecular markers have been used in the genetic characterization of oats, including, including Amplified Fragment Length Polymorphism-AFLP (Achleitner et al., 2008), microsatellites or Simple Sequence Repeats-SSR (Li et al., 2000; Wight et al., 2010; Tanhuanpää et al., 2012), and Single Nucleotide Polymorphism-SNP (Tinker et al., 2014; Zimmer et al., 2019), among others. Despite the rapid development of new molecular tools, SSRs continue to be used, and are considered suitable for genetic diversity studies, and for the inference of population structure (Rana et al., 2019; Raza et al., 2020; Arora et al., 2021). SSRs are also effective in detecting heterogeneity in oat varieties, purity of seed lots, and genetic mapping and fingerprinting studies (Wight et al., 2010; Zheng et al., 2019). SSRs are not affected by external and internal environments, are cost-effective, fast, accurate, simple, highly polymorphic, reliable, and their co-dominant inheritance allow them to distinguish between homozygous and heterozygous loci (Jannink and Gardner, 2005; Zheng et al., 2019; Kaur et al., 2021).

For these reasons, the present study focusses on the phenotypic and SSR-based genetic analysis of 132 oat genotypes of diverse origin, including historical and current advanced pure lines and cultivars. The assessment of the phenotypic and genetic diversity, with other important parameters for the estimation of genetic gain, were addressed with the purpose of understanding the possible

causes of the low rate of cultivar turnover in Chile, as well as the status of the available germplasm regarding the breeding goals in the selection of conventional food-oats.

2 Materials and methods

2.1 Plant material

A collection of 132 oat genotypes of diverse origin representing 85 pure lines, 46 cultivars, and one land race, including historical and modern germplasm, was selected for the study (Supplementary Table S2). Seeds of foreign germplasm were provided by the Quaker International Oat Nursery-QION and International Oat Nursery-ION collaborations, and other breeding programs through specific Material Transfer Agreements. The seeds of cultivars and lines created and/or registered in Chile were obtained from the INIA breeding program. The geographic origins of introduced lines were obtained from the POOL online oat database (Tinker and Deyl, 2005).

2.2 Field trial and experimental design

A field trial including the 132 oat genotypes was sown on June 15, 2020, in Vilcún (38°41′25′′S, 72°23′32′′ W, La Araucanía Region, Chile). The experimental design was an alpha-lattice with two replications. The experimental unit was a 2 m long and 0.6 m wide plot, resulting in a total of 264 experimental plots. The experiment was arranged in 12 incomplete blocks, each block containing 11 experimental units each corresponding to a different oat genotype. The seeds were disinfected using 2 mg Benomyl (Polyben 50 WP, Anasac S.A., Chile) per g of seeds; the seeding rate was 12 $g \cdot m^{-2}$. The agronomic management consisted of standard control of weeds and fertilizer application, without insecticide or fungicide treatment. The nutrient rates were 14 $g \cdot m^{-2}$ N, using 27% magnesium calcium ammonium nitrate applied 20% at seeding, 40% at early tillering (Zadoks Stage Z-21), and 40% during full tillering (Z-27) (Zadoks et al., 1974); 8 $g \cdot m^{-2}$ P_2O_5 applied as monocalcium phosphate at seeding; and 6 $g \cdot m^{-2}$ K_2O using potassium/magnesium sulfate mixed with monocalcium phosphate at seeding. The herbicide doses were 0.08 $mL \cdot m^{-2}$ of a mix of Flufenacet, Flurtamona, and Diflufenican (Baccara Forte 360 SC, Bayer AG, Leverkusen, Germany) at crop pre-emergence, 0.08 $mL \cdot m^{-2}$ S-metolacloro (Dual Gold 960 EC, Syngenta S.A., Cartagena, Colombia) at early crop emergence, and 0.07 $mL \cdot m^{-2}$ MCPA-dimethylammonium (MCPA 750 SL, A.H. Marks Co., West Yorkshire, England) at full tillage.

2.3 Phenotypic trait measurements

A total of 28 phenotypic traits were evaluated in each experimental unit of the field trial (Supplementary Table S1). Phenotypic traits were selected based on USDA online Triticeae Toolbox Oat - T3 Oat (https://oat.triticeaetoolbox.org/search/traits), being mostly quantitative traits of importance for breeding. Visual scores of diseases and agronomic types were assessed during tillage, panicle emission, dough and mature grain stages. Plant height and panicle length were measured with a ruler at maturity; lodging was evaluated one day before harvest. After harvest, the grain yield was normalized to 12% moisture. Subsequently, a clean representative sample of 250 g of each experimental unit was obtained for evaluation of quality traits in hulled and dehulled grains.

2.4 DNA isolation and SSRs analysis

Ten seeds of each oat genotype were germinated and grown in Petri dishes containing paper towel at room temperature, being periodically moistened with distilled water. Then, 10 mg of fresh leaf tissue were collected from each of the 10 seedlings and pooled, resulting in a total of 100 mg fresh sample per oat genotype. Each pooled sample was fully pulverized using liquid nitrogen. DNA isolation was conducted following a CTAB DNA isolation procedure (Fulton et al., 1995). In previous work conducted in our laboratory in 2014, we had selected 107 SSRs reported in published literature based on their reported quality. Primer pairs for the selected SSRs were then synthesized by Integrated DNA Technologies Inc., USA. A group of 38 SSRs were subsequently pre-selected based on their reproducibility and quality tested against several oat genotypes available at the time under our local laboratory conditions (unpublished work). In the present study, the 38 SSRs previously selected were screened against a subsample of 10 oat genotypes out of the 132 total genotypes studied herein. The selected genotypes were representative of different pedigrees, geographic origins, and included modern and historic cultivars, and pure lines. Out of the 38 SSRs tested, 14 SSRs were selected for the present evaluation of 132 oat genotypes based on their polymorphism and visual quality of the amplification patterns. The polymerase chain reactions (PCR) were conducted in a final volume of 12 µL, using 20 ng DNA, 1X PCR buffer, 0.125 mM dNTPs, 1.5 mM $MgCl_2$, 2.5 mM of each primer and 1 U Taq polymerase. The PCR program included a first denaturing step of 94°C for 7 min, 40 cycles of 95°C for 1 min, 54-58°C for 45 sec and 72°C for 1 min, followed by a final elongation at 72°C for 7 min. The amplified fragments were separated in 6% polyacrylamide sequencing gels, at 80 W for 2 to 3 hours, depending on the marker. The fragments were visualized with silver nitrate stain (Narváez et al., 2001).

2.5 Phenotypic data analysis

The variance analysis of the phenotypic data was carried out using the model ($Y_{ijk} = \mu + Gen_i + Block(Rep_j) + \varepsilon_{ijk}$), which was fitted by applying a restricted maximum likelihood- REML, using the R 4.1.0 software (R Core Team, 2022) along with the Metan package (Olivoto and Lúcio, 2020). The genotype was treated as a random effect to estimate the broad-sense heritability (H^2) of the 28 phenotypic traits, and the Best Linear Unbiased Predictor-BLUP of

the 132 oat genotypes. A multi-trait genotype-ideotype distance index (MGIDI) was calculated for each genotype, using 28 phenotypic traits with their respective breeding objectives (increase, decrease) (Supplementary Table S1) (Olivoto and Nardino, 2021).

The linear associations between traits were examined by Pearson correlations, using the adjusted BLUP means obtained through the model. To identify the linear effect between the traits, controlling statistically the effect of others traits, a Pearson correlation matrix was used to calculate partial correlations (Olivoto and Lúcio, 2020). The significance of 465 pairwise trait-trait combinations was tested with the Student's t ($P \leq 0.05$). The correlation coefficients were ranked as negligible ($|r| < 0.10$), weak ($0.10 \geq |r| \leq 0.39$), moderate ($0.40 \geq |r| \leq 0.69$), strong ($0.70 \geq |r| \leq 0.89$), and very strong ($|r| \geq 0.90$) (Mukaka, 2012).

The phenotypic diversity of each trait was estimated with the Shannon-Weaver diversity index (H) (Perry and McIntosh, 1991), using the formula: $H = -\sum_{i=1}^{n} P_i \ log_2 \ P_i$, where n is the number of classes of traits and P_i is the proportion of accessions in the $_{ith}$ class of a trait. Each H value was normalized by dividing it by its maximum value (log2n), which ensured that all values were in the range of 0 to 1 for comparison purposes, called H' (Perry and McIntosh, 1991). Since the traits were mainly quantitative, the BLUP data were previously scaled to percentages and divided into ten different phenotypic classes ranging from 0 to 100%. The diversity index was categorized as high ($H' \geq 0.60$), intermediate ($0.40 \geq H' < 0.60$) or low ($H' < 0.40$) (Yemataw et al., 2018).

To analyze grouping patterns of the oat germplasm, a principal component analysis – PCA and a cluster analysis using the resulting PC coordinates were conducted using the adjusted BLUP means of the 28 phenotypic traits. The analysis was performed using Factoextra (Kassambara and Mundt, 2020), and FactoMineR (Lê et al., 2008) R packages. A group of six oat genotypes were used as references in figures.

2.6 Genetic data analyses

The SSR alleles were scored as presence/absence (0/1) and recorded in a data matrix. The polymorphism information content (PIC) was calculated for each marker with the formula: $PIC = 1 - \sum_{i=1}^{n} Pi^2$, where n is the number of alleles, and *Pi* is the frequency of allele *i* (Weir, 1996). The discriminating power (D) of each marker was estimated with the online iMEC software (Amiryousefi et al., 2018). A genetic distance tree was generated with the Jaccard dissimilarity index, and the unweighted neighbor joining clustering method with 10,000 bootstrap reps, using Darwin 6.0.21 (software available in https://darwin.cirad.fr/).

The genetic structure of the oat germplasm was inferred with the *Structure 2.3.4* software (Pritchard et al., 2000). The simulations assumed K values from 1 to 10 populations, 100,000 burn-in run iterations, 100,000 Markov Chain Monte Carlo, with 10 runs for each K value. The optimal number of populations was estimated observing the Delta K decay with each K value (Evanno et al., 2005). Then the Shannon Information diversity Index (I), Nei's genetic diversity index (He), observed heterozygosity (Ho), number of alleles, and private alleles, were calculated using GeneAlex v6 (Peakall and Smouse, 2006). Low frequency alleles were pooled to fulfill the data format of GeneAlex. A molecular variance analysis (AMOVA) was also carried out using GeneAlex.

Finally, the phenotypic and genetic diversity indexes, and the MGIDI selection index were calculated in the phenotypic clusters and genetic pools inferred, and in categorical groups formed by origin of the germplasm. Then Kruskal-Wallis non-parametric comparisons were carried out between the groups, with $P \leq 0.05$ declared as significance and $0.05 < P \leq 0.10$ considered as a tendency.

3 Results

3.1 Phenotypic traits variation and heritability

The effect of genotype on phenotype was significant over all traits ($P \leq 0.05$), showing variation in the studied oat germplasm in all the traits (Table 1). The genetic coefficient of variation (CVg) ranged from 0.22% (dry matter content) to 160% (lodging percentage); it was low for grain quality traits such as hectoliter weight (4.16%), groat content (5.65%), dry matter (0.22%) and heading days (3.41%) (Table 1). High CVg were found in the incidence of diseases such as *Pseudomonas syringae* (76.96%) and Barley Yellow Dwarf Virus (98.68%), and in the low severity (66.21%) and high severity (62.97%) groat staining (Table 1). The lodging percentage was the only trait exhibiting CVg above 100%, showing the high variation in the data (0.4% - 0.94%) in relation to the mean (15.39%) (result not shown). This is likely due to the field trial having low lodging except for a group of sensitive genotypes exhibiting severe lodging (Supplementary Table S2).

The contribution of genetic variance to the phenotypic variance ranged from 15.62% (plant height at tillering) to 98.42% (panicle type), resulting in broad sense heritability (H^2) ranging from 0.16 to 0.98 (Table 1). The H^2 was low for grain yield (0.27) and traits measured at early stages of development in the field, such as vigor score (0.19), agronomic score at tillering (0.34), plant height at tillering (0.16) and *P. syringae* incidence (0.23), showing a major proportion of environmental factors explaining the phenotypic variance. In contrast, high H^2 was observed for grain quality traits such as hectoliter weight (0.82), groat content (0.81), groat protein (0.78) and fat (0.78) content, and morphological traits like panicle type (0.98) and hull color (0.84).

3.2 Phenotypic trait diversity

To identify traits as potential breeding targets in the studied germplasm, the phenotypic diversity was estimated based on the Shannon-Weaver diversity index (H). H was normalized by its maximum number of classes, obtaining H' values in the range between 0 and 1 for comparative purposes (Perry and McIntosh, 1991). Thus H' allowed us to classify the traits in different diversity categories, excluding the effect of the number of classes. The H

TABLE 1 Effect of the genotype on phenotypic traits and genetic parameters.

Traits	Variance				*P* value	H^2	CVg (%)
	Genetic	Rep:block	Residual	Phenotypic			
Agronomic							
Grain yield	13,754	9,005	28,638	51,397	< 0.001	0.27	13.74
Heading days	29.00	0.15	2.60	31.76	< 0.001	0.91	3.41
Lodging percentage	613.10	5.45	146.48	765.10	< 0.001	0.80	160.90
Lodging severity	0.40	0.02	0.60	1.08	< 0.001	0.43	33.42
Panicle length	6.60	0.00	3.01	9.67	< 0.001	0.69	12.26
Plant height at tillering	8.00	9.16	34.55	51.81	0.048	0.16	4.22
Plant height at maturity	151.10	33.55	33.57	218.30	< 0.001	0.69	9.60
Plant types							
Vigor score	0.05	0.03	0.19	0.26	0.021	0.19	7.99
Agronomic score at tillering	0.17	0.06	0.27	0.50	< 0.001	0.34	6.39
Agronomic score at dough grain	0.94	0.03	0.45	1.42	< 0.001	0.66	15.35
Agronomic score at maturity	0.82	0.00	0.33	1.15	< 0.001	0.71	14.36
Hull color	0.70	0.00	0.13	0.83	< 0.001	0.84	50.79
Panicle type	0.47	0.00	0.01	0.48	< 0.001	0.98	29.03
Incidence of diseases							
Barley Yellow Dwarf Virus	0.15	0.04	0.28	0.48	< 0.001	0.31	98.68
Drechslera avenae	187.75	26.01	148.91	362.67	< 0.001	0.52	53.71
Pseudomonas syringae	7.11	1.47	22.61	31.20	0.007	0.23	76.96
Grain quality							
Hectoliter weight	5.14	0.20	0.90	6.24	< 0.001	0.82	4.16
Groat content	14.79	0.50	3.06	18.34	< 0.001	0.81	5.65
Broken groats after peeling	12.35	0.00	3.00	15.35	< 0.001	0.80	38.99
Hulled grains after peeling	0.80	0.00	0.50	1.29	< 0.001	0.61	55.41
Hull staining	0.10	0.02	0.20	0.32	< 0.001	0.32	22.91
Low severity groat staining	17.54	0.56	4.57	22.66	< 0.001	0.77	66.21
High severity groat staining	0.44	0.01	0.38	0.83	< 0.001	0.53	62.97
Thousand hulled grain weight	14.75	1.17	2.66	18.58	< 0.001	0.79	8.53
Thousand dehulled grain weight	11.27	0.39	1.26	12.92	< 0.001	0.87	10.51
Groat protein	3.10	0.46	0.44	3.99	< 0.001	0.78	11.40
Groat fat	0.52	0.01	0.13	0.66	< 0.001	0.78	9.11
Groat dry matter	0.04	0.02	0.02	0.08	< 0.001	0.51	0.22

Rep, replicate; H^2, broad sense heritability; CVg, genetic coefficient of variation.

index, reflecting the abundance of oat genotypes in different phenotypic classes, was 1.71 ± 0.69, ranging from 0.45 (hectoliter weight) to 2.63 (*Drechslera avenae* incidence); whereas H' was 0.68 ± 0.18, ranging between 0.28 (hectoliter weight) and 0.94 (lodging severity) (Table 2). Twenty-four of 28 traits were in the high (H' > 0.6), four traits in the intermediate (0.40 ≥ H'< 0.60) and two traits in the low (H'< 0.40) diversity categories.

Grain yield had high diversity (H' = 0.73). However, only 14% of the oat genotypes outperformed Supernova INIA (Table 2). Most of the genotypes were in the high phenotypic classes for grain quality traits such as hectoliter weight, groat content and thousand hulled grain weight, but a few genotypes had higher quality than Supernova INIA. A similar pattern was observed for heading days with most genotypes exhibiting intermediate to long cycles, but 2%

10.3389/fpls.2023.1298591

TABLE 2 Frequency of the germplasm in different phenotypic classes and phenotypic diversity.

Traits	BLUPs range		Frequency at each phenotypic class (%)										Diversity	
	Min	Max	1	2	3	4	5	6	7	8	9	10	H	H'
Agronomic														
Grain yield, g · m^{-2}	596.5	1,035.4	–	–	–	–	–	2	6	24	55	14	1.70	0.73
Heading days, d	138.4	177.9	–	–	–	–	–	–	–	2	60	38	1.10	0.69
Lodging percentage, %	0.4	91.4	67	11	2	6	2	2	1	2	3	4	1.80	0.54
Lodging severity, 1-5 rating	1.3	3.1	–	–	–	–	30	13	15	12	23	7	2.43	0.94
Panicle length, cm	15.7	29.9	–	–	–	–	–	8	40	44	5	2	1.70	0.73
Plant height at tillering, cm	64.4	73.4	–	–	–	–	–	–	–	–	16	84	0.63	0.63
Plant height at maturity, cm	97.2	166.3	–	–	–	–	–	2	12	58	24	5	1.62	0.70
Plant types														
Vigor score, 1-4 rating	2.2	3.0	–	–	–	–	–	–	–	2	30	68	0.99	0.62
Agronomic score at tillering, 1-10 rating	5.1	6.9	–	–	–	–	–	–	–	2	23	76	0.88	0.56
Agronomic Score at dough grain, 1-10 rating	2.8	7.7	–	–	–	1	2	3	7	30	39	20	2.07	0.74
Agronomic score at maturity, 1-10 rating	3.5	7.7	–	–	–	–	2	5	7	29	45	14	1.99	0.77
Hull color, 0-5 rating	0.1	4.7	1	–	50	5	30	2	6	1	3	2	1.97	0.62
Panicle type, 1-3 rating	1.0	2.9	–	–	–	12	–	–	39	–	2	47	1.50	0.75
Incidence of diseases														
Barley Yellow Dwarf Virus, %	0.0	1.6	12	41	23	8	9	5	1	1	1	1	2.40	0.72
Drechslera avenae, %	7.4	60.4	–	6	22	20	24	13	2	6	5	2	2.73	0.86
Pseudomonas syringae, %	1.7	11.4	–	30	33	21	8	4	2	2	–	1	2.28	0.76
Grain quality														
Hectoliter weight, kg · hL^{-1}	45.9	64.5	–	–	–	–	–	–	–	8	92	1	0.45	0.28
Groat content, %	53.6	96.5	–	–	–	–	–	1	43	55	–	1	1.10	0.55
Broken groats after peeling, %	1.1	23.5	1	3	23	36	24	6	4	1	1	2	2.34	0.70
Hulled grains after peeling, %	0.3	4.9	1	17	41	20	10	5	2	2	2	1	2.39	0.72
Hull staining, 0-3 rating	0.7	1.9	–	–	–	1	–	8	37	33	16	6	2.06	0.80
Low severity groat staining, %	1.3	20.7	3	32	26	20	5	5	4	2	3	1	2.55	0.77
High severity groat staining, %	0.3	3.5	2	31	27	19	9	7	4	1	–	1	2.45	0.77
Thousand hulled grain weight, g	28.1	55.3	–	–	–	–	–	1	3	36	55	5	1.41	0.61
Thousand dehulled grain weight, g	23.4	39.5	–	–	–	–	–	1	11	31	43	14	1.85	0.80
Groat protein, %	11.7	21.3	–	–	–	–	–	5	40	39	15	2	1.77	0.76
Groat fat, %	6.0	9.3	–	–	–	–	–	–	2	26	45	27	1.65	0.83
Groat dry matter, %	89.9	90.9	–	–	–	–	–	–	–	–	–	100	0.00	0.00
Overall													1.71	0.68
SD													0.69	0.18

BLUPs, best linear unbiased predictors; H, Shannon Weaver diversity index; H', H scaled by the number of phenotypic classes. The phenotypic diversity was categorized as high (H' ≥ 0.60), intermediate (0.40 ≥ H'< 0.60) or low (H'< 0.40). The underlined numbers indicate the classes of the reference cultivar Supernova INIA.

of genotypes emitted their panicles earlier than Supernova INIA. Plant height at maturity exhibited high diversity (H' = 0.70), with a high proportion of intermediate to tall genotypes.

Groat protein (H' = 0.76) and fat (H' = 0.83) content showed high diversity, with a similar fraction of genotypes with higher and lower values than Supernova INIA (Table 2). High diversity was

also found for grain quality traits including hulled groats and broken grains after peeling, high and low severity grain staining, hull staining, and foliar diseases; a good proportion of genotypes had better quality and tolerance to diseases than Supernova INIA. All oat genotypes were assigned to the same phenotypic class for groat dry matter content, resulting in null diversity.

3.3 Pairwise correlations between traits

Since Pearson´s correlation does not consider the influence of other traits on the relationship between two traits, a partial correlation analysis was used to control statistically the effect of other traits on the correlations (Olivoto and Lúcio, 2020). Forty-four associations were significant both with Pearson and partial correlation analysis, representing robust associations (Figure 1). However, 64 associations were significant only with Pearson correlation, whereas 18 associations were found significant only in the partial correlation analysis (Supplementary Table S3).

As expected, the association between related traits such as thousand hulled and dehulled grain weight and agronomic scores at dough and mature grain stages was strong, whereas the associations between the other traits were mostly weak to moderate (Figure 1). Grain yield exhibited positive associations with the agronomic scores at tillering and maturity, and negative associations with lodging percentage, plant height at maturity and groat protein content. Therefore, grain yield associations were mostly favorable for genetic breeding, except for groat protein. Also, plant height at maturity exhibited a positive association with lodging percentage, plant height at tillering and panicle length.

Hectoliter weight showed a positive association with groat content and panicle type, and negative associations with hull color and hull staining; all these associations are favorable for genetic breeding of high grain quality oats (Figure 1). As

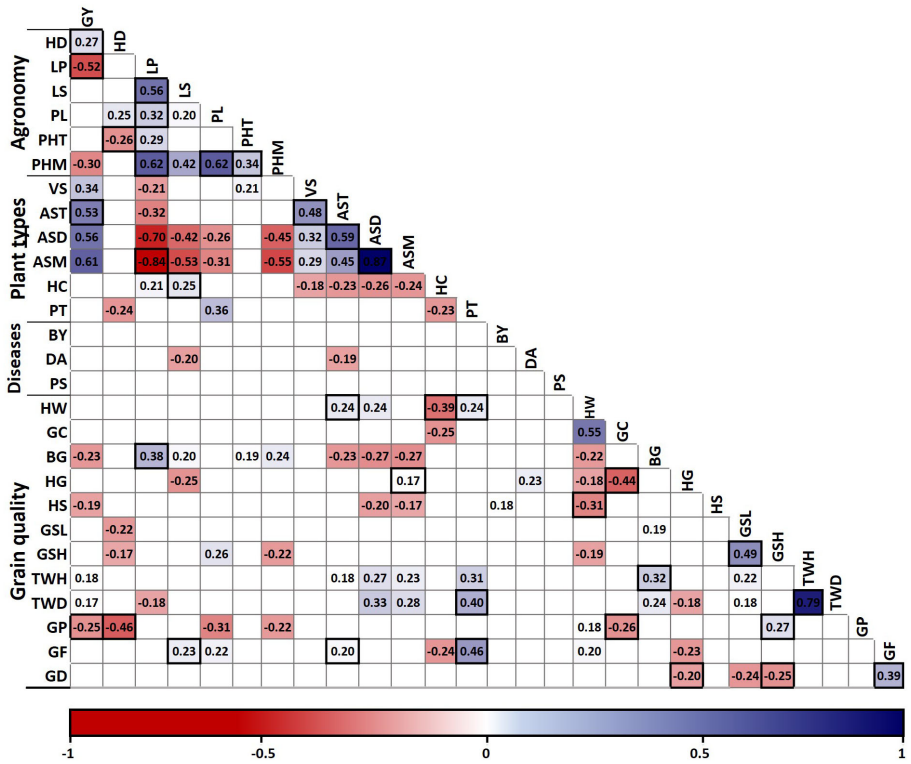


FIGURE 1
Pearson correlation matrix of 28 traits measured in 132 oat genotypes and partial correlation validation. Only correlations with $P \leq 0.05$ are shown in the graph. Significant associations in the partial correlation analysis are marked with black boxes. GY, grain yield; HD, heading days; LP, lodging percentage; LS, lodging severity; PL, panicle length; PHT, height at tillering; PHM, height at maturity; VS, vigor score; AST, agronomic score at tillering; ASD, dough grain; and ASM, maturity; HC, hull color; PT, panicle type; BY, barley yellow dwarf virus incidence; DA, *Drechslera avenae* incidence; PS, *Pseudomonas syringae* incidence; HW, hectoliter weight; GC, groat content; BG, broken groats and HG, hulled grains after peeling; HS, hull staining; GSL, low severity and GSH, high severity groat staining; TWH, thousand hulled and TWD, thousand dehulled grain weight; GP, groat protein; GF, groat fat; and GD, groat dry matter content.

expected, high and low severity groat staining exhibited a positive association. However, high severity groat staining had a positive association with protein content and negative with groat dry matter content, whereas low severity groat staining exhibited a negative association with heading days. Also, groat protein had a negative association with heading days and groat content.

3.4 Phenotypic principal component analysis and clustering of the oat germplasm

A multivariate analysis was performed on the 132 oat genotypes values, to reduce complexity and explore the relationships among several traits of economic importance. Dimension 1 (17.41%) was mainly represented by agronomic plant type scores and grain yield, while Dimension 2 (11.09%) was associated with grain quality traits (Supplementary Figures S1A, B). The genotypes resulted in a mostly well distributed germplasm but with a group of nine genotypes markedly separated from the others (Supplementary Figure S1C).

The grouping analysis of the principal component coordinates found significant explanation of the clusters mainly in 13 traits linked to the first, second and fourth PCA dimensions, representing 35.17% of the explained variance (Figure 2A; Supplementary Table S4). Depending on the correlations of the traits in relation to the dimensions of the PCA (Figure 2B), each cluster exhibited a higher or lower mean compared to the overall mean of the 132 oat genotypes (Supplementary Table S5).

The oat genotypes were grouped in three phenotypic clusters, Cluster 1 (N = 10), Cluster 2 (N = 63) and Cluster 3 (N = 59)

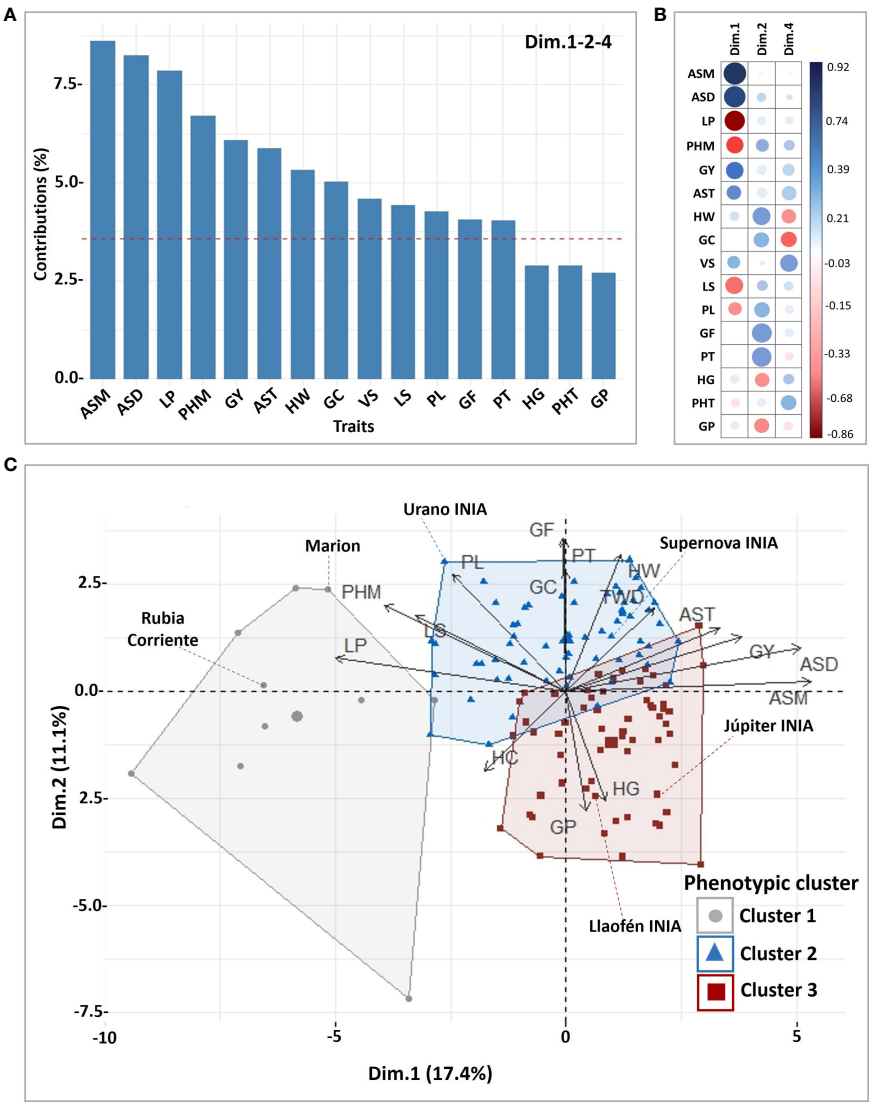


FIGURE 2
Phenotypic cluster analysis for 28 traits and 132 oat genotypes. **(A)** Contribution of the main traits explaining the variance of significant ($P \leq 0.05$) dimensions in the cluster analysis, **(B)** correlation of the traits and the dimensions, **(C)** biplot of genotypes and traits showing the resulting clusters. GY, grain yield; LP, lodging percentage; LS, lodging severity; PL, panicle length; PHT, height at tillering; PHM, height at maturity; VS, vigor score; AST, agronomic score at tillering; ASD, dough grain; and ASM, maturity; PT, panicle type; HW, hectoliter weight; GC, groat content and GF, groat fat; HG, hulled grains; GP, groat protein.

(Figure 2C; Supplementary Table S5). Cluster 1 had long panicles but with very high plant height and high lodging, low grain yield, and low industrial grain quality. It was composed mainly of foreign historical cultivars from United States of America (N = 3), Canada (N = 1), Italy (N = 1), Austria (N = 1), Australia (N = 1), Uruguay (N = 1), a pure line from Brazil (Supplementary Table S2), and a Chilean landrace (Rubia Corriente), widely used in Chile as forage. Cluster 2 had high grain yield, good industrial grain quality, intermediate plant height, intermediate panicle length, moderate to low groat protein, and slightly high groat fat, such as Supernova INIA and Urano INIA. Cluster 3 genotypes had overall lower plant height, lower lodging, high groat protein, slightly low fat groat contents, slightly higher groat staining, lower industrial grain quality, and shorter panicles, than Cluster 2, for example Júpiter INIA.

3.5 Phenotypic value of the oat germplasm for oat-food breeding

To estimate the phenotypic value of the oat germplasm we estimated a Multi-trait Genotype-Ideotype Distance Index (MGIDI) that quantifies each oat genotype with regard to breeding objectives; the lower the index the closer the individual is to the ideotype and therefore higher genetic gain is expected (Olivoto and Nardino, 2021). The MGIDI considered the 28 traits with their respective breeding objectives (increase, decrease), according to the expected attributes in conventional food-oats, such as high field performance and grain quality (Supplementary Table S1).

The MGIDI index ranged from 3.46 to 8.91 with a mean of 5.66, while the reference cultivar Supernova INIA had a value of 4.78 (Figure 3A, Supplementary Table S2). A group of 26 genotypes with better phenotypic performance than Supernova INIA (MGIDI< 4.78) were selected as promising material for food-oat commercial breeding. The group had selection gain in almost all traits, excepting plant height at maturity and groat fat content (Supplementary Table S6). The selected genotypes showed different qualities in relation to Supernova INIA (Figure 3B); the germplasm was mainly from Chile (N = 21), Canada (N = 2), United States of America (N = 2) and New Zealand (N = 2) (Supplementary Table S2). The genotypes ranked in the lower 5% extreme of the MGIDI scores were chosen as potential candidate lines for direct development of new food-oat cultivars. The 105 genotypes with higher MGIDI (lower phenotypic performance) scores than Supernova INIA were kept for pre-breeding purposes, due to their specific characteristics of interest and possibly other potential non-characterized beneficial properties, their conservation being relevant for future studies.

3.6 Genetic variation revealed by the SSRs

A total of 64 alleles from 14 SSRs were detected in the 132 oat genotypes. These markers were polymorphic, and 20 alleles exhibited frequencies lower than 0.05 in the set of samples (Table 3). The number of alleles per marker ranged from 2 (AME019, AME055, and AME076) to 10 (AME178), with an average of 4.27 alleles per locus; the PIC ranged between 0.26 (AME177) and 0.84 (AME178), with an average of 0.58; and the discriminatory power ranged from 0.28 (MAMA_5) to 0.57 (AME102), with a mean of 0.43. Thus, the

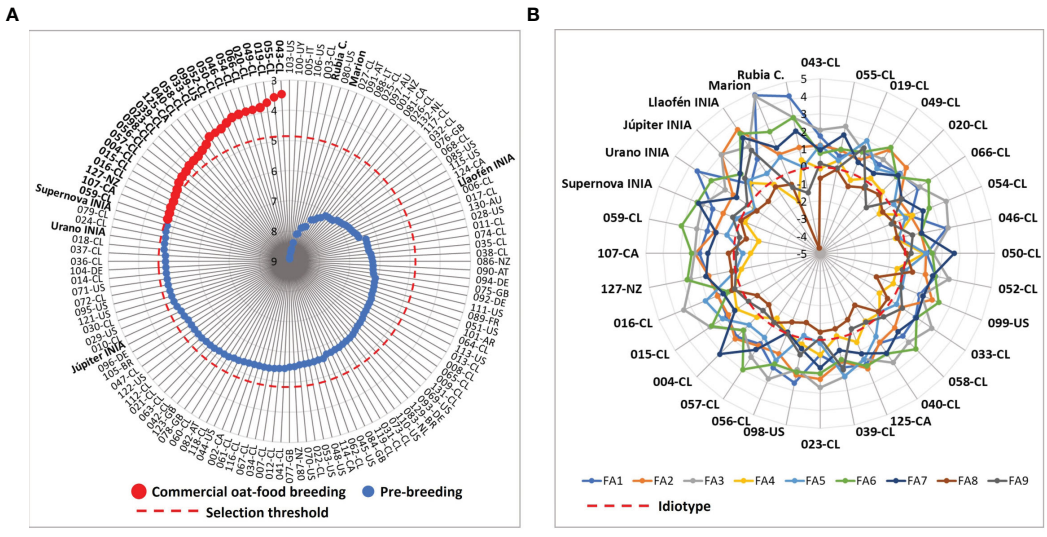


FIGURE 3
Multi-trait genotype–ideotype distance indexes (MGIDI) of the 132 oat genotypes considering the objectives for conventional food-oat breeding with 28 traits. **(A)** selection of the better germplasm for commercial food-oat breeding and pre-breeding, according to MGIDI scores, **(B)** weaknesses and strengths, corresponding to the qualities for breeding of the selected genotypes and reference cultivars. FA: factors. FA1: grain yield, height at maturity, panicle length, lodging percentage, lodging severity, dough grain and maturity; FA2: thousand hulled and dehulled grains weight, broken grains after peeling; FA3: low severity groat staining, high severity groat staining, *Pseudomonas syringae* incidence; FA4: plant height at tillering, vigor score, agronomic score at tillering; FA5: groat content, hulled grains after peeling, *Drechslera avenae* incidence; FA6: groat protein, heading days; FA7: hull staining, barley yellow dwarf virus incidence; FA8: hectoliter weight, hull color; FA9: groat dry matter, groat fat content, panicle type.

10.3389/fpls.2023.1298591

TABLE 3 Primer sequences, and efficiency indexes of the SSRs.

SSRs	5´-> 3´Primer sequences	SSR Type	Ta (°C)	Alleles (N)			PIC	D	SSR´s reference
				Common (> 0.05)	Rare (≤ 0.05)	Total			
AM14	F: TGGGTGGCGAAGCGAATC	Genomic	58	4	0	4	0.33	0.44	Li et al. (2000)
	R: GTGGTGGGCACGGTATCA								
AM22	F: AAGAGCGACCCAGTTGTATG	Genomic	56	2	1	3	0.61	0.49	Li et al. (2000)
	R: ATTGTATTTGTAGCCCCAGTTC								
AME013	F: ACGGAACTTCAACACTTTGG	EST	56	2	2	4	0.55	0.38	Tanhuanpää et al. (2012)
	R: GGCATGAGAGTTTTTATGAACC								
AME019	F: CATCACAGTCGCAGCCATG	EST	58	2	0	2	0.56	0.38	Tanhuanpää et al. (2012)
	R: GCATGCATTTTCCCCTCACG								
AME055	F: TTCGACCATGGGAATCTTTG	EST	56	2	0	2	0.57	0.54	Tanhuanpää et al. (2012)
	R: CGGAGGTGCAAACCCTAGTA								
AME076	F: CATGATCCATCACACATACCG	EST	56	2	0	2	0.57	0.57	Tanhuanpää et al. (2012)
	R: CGAATGGATGCTGAATTGG								
AME102	F: GCTGCCTCTACATGAGCAGA	EST	58	4	0	4	0.70	0.57	Tanhuanpää et al. (2012)
	R: TCCTCCTCCAGGATGTGACT								
AME154	F: GTACACACATCCAATCCATTTC	EST	54	3	0	3	0.69	0.46	Tanhuanpää et al. (2012)
	R: TGAAGGAACGGAAATCTGAAG								
AME177	F: ATCGGGTACTAGTGATACATAC	EST	54	3	3	6	0.26	0.54	Tanhuanpää et al. (2012)
	R: CATGTATCTCATCCCAAACTC								
AME178	F: TGTCTTATCTGGCTGGAGCA	EST	56	4	6	10	0.84	0.43	Tanhuanpää et al. (2012)
	R: AGAATTGGAACCGTGTGTAAC								
MAMA_1	F: CATGCTGGCGAAATCTATCA	Genomic	56	5	0	5	0.74	0.47	Wight et al. (2010)
	R: GTGCGCCTCTAACGAAAAAT								
MAMA_5	F: AACCCTAATTACTGCTCCGTTTC	Genomic	56	4	4	8	0.79	0.28	Wight et al. (2010)
	R: GGATTGGGACTTCGCATCTA								
MAMA_11	F: GACTACCGCCCAGATGAGAC	Genomic	56	3	4	7	0.67	0.51	Wight et al. (2010)
	R: TGTATGCACCGATGCAATTT								
MAMA_13	F: CGATGCACTCAGATTTGGAA	Genomic	56	4	0	4	0.79	0.33	Wight et al. (2010)
	R: CTGGATCAAGCAGACATGGA								
Mean				2.80	1.30		0.58	0.43	
Total				44	20				

F, forward; R, reverse; Ta, annealing temperature of primers; PIC, polymorphism information content; D, discriminating power.

polymorphism and discrimination power achieved by this set of SSRs can be considered intermediate. The allele frequencies are detailed in Supplementary Figure S2.

Two main groups were visually different in the neighbor joining genetic tree; sub-tree I grouped mainly Chilean germplasm with a low proportion of introduced lines and cultivars such as Supernova INIA (126-NZ) and Urano INIA (128-CA); sub-tree II grouped mostly foreign germplasm, historical germplasm, and a low proportion of Chilean advanced lines (Figure 4A; Supplementary Table S2). The genetic dissimilarity between the oats was high, with an average of 0.75, indicating a high degree of genetic differentiation between the oat genotypes, with exception of the Chilean pure lines 59-CL and 60-CL, revealed as duplicates (Figure 4B; Supplementary Table S2).

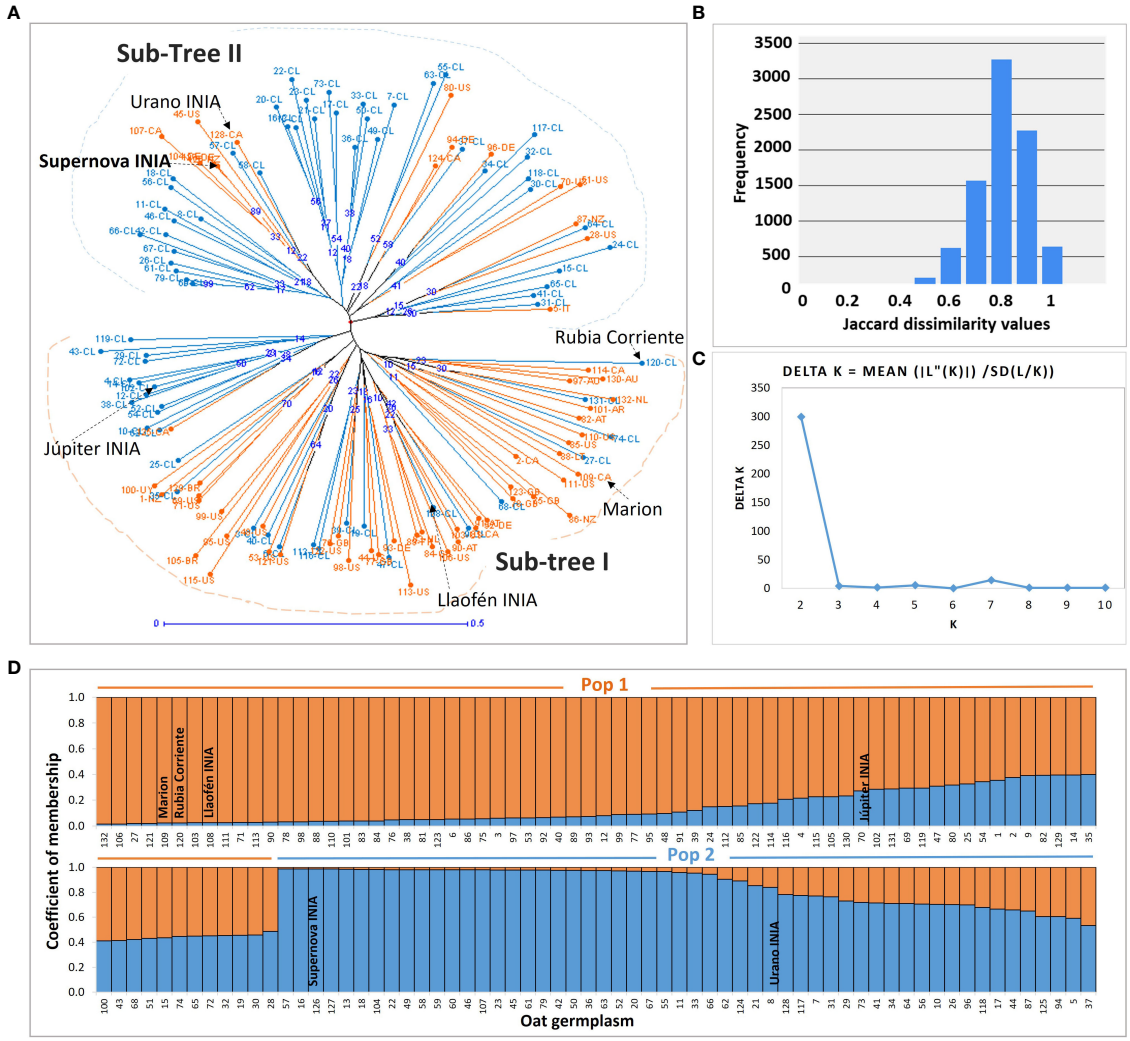


FIGURE 4
Genetic tree and population structure of the oat germplasm. **(A)** The genotypes created in Chile and the foreign germplasm are marked in light blue and orange, respectively; blue numbers in nodes are the bootstrap support of the branches; sub-tree I, and sub-tree II, are the main groups detected visually. **(B)** Frequency histogram of the Jaccard dissimilarity indexes with 132 oats and 10,000 bootstraps. **(C)** Change in likelihood of the data L(K) at values of K populations from 1 to 10, used to infer the true value of (K) **(D)** Populations inferred with the Structure algorithm. The number labels in **(A, D)** representing the 132 oat genotypes are consistent with Supplementary Table S2.

3.7 Population structure

The Structure analysis with the SSRs supported two subpopulations based on the decay of Delta K, suggesting the existence of two genetic pools in the germplasm (Figures 4C, D). The pairwise genetic distance and the similarity between Pop 1 and Pop 2, were 0.24 and 0.79, respectively (Supplementary Table S7). The molecular variance analysis found significant fixation indexes, indicating significant genetic variation among subpopulations (F_{ST}), among individuals (F_{IS}), and within individuals (F_{IT}) (P = 0.001), representing the 11%, 77%, and 12% of the molecular variance, respectively (Supplementary Table S8).

Fifty-nine percent (N = 78) of the genotypes were assigned to Pop 1, and the other 41% (N = 54) to Pop 2. However, 23.48% of the germplasm was admixed, corresponding to genotypes exhibiting less than 0.7 membership to their respective subpopulation (Wang et al., 2023). The admixed germplasm was higher in Pop 1 (N = 23) than in Pop 2 (N = 8) (Supplementary Table S2). Most foreign and historical genotypes, such as Rubia Corriente, Eva, Llaofén INIA, also of the cultivar Júpiter INIA, belonged to Pop 1 (Supplementary Table S2). Seventy percent of modern Chilean advanced pure lines, and the most important commercial cultivars in Chile such as Supernova INIA and Urano INIA, were assigned to Pop 2.

The comparison of the genetic distance tree to the population structure showed a similar but not identical grouping of the genotypes; sub-tree I was analogous to Pop 1, and sub-tree II was similar to Pop 2. Six and nine genotypes of sub-tree I and sub-tree- II, respectively, were assessed inversely to Pop 2 and Pop 1 (Supplementary Table S2).

TABLE 4 Genetic diversity indexes in the inferred subpopulations and in all genotypes.

SSRs	Sub-population 1 (N = 78)					Sub-population 2 (N = 54)					Total (N = 132)				
	Na	Ne	I	Ho	He	Na	Ne	I	Ho	He	Na	Ne	I	Ho	He
AM14	4	2.01	0.95	0.35	0.50	4	3.26	1.25	0.67	0.69	4	2.61	1.15	0.48	0.62
AM22	3	1.81	0.76	0.00	0.45	2	1.05	0.12	0.00	0.05	3	1.47	0.58	0.00	0.32
AME013	4	2.23	0.95	0.04	0.55	4	1.35	0.53	0.00	0.26	4	1.86	0.83	0.03	0.46
AME019	2	1.60	0.56	0.00	0.38	2	1.86	0.65	0.00	0.46	2	1.71	0.61	0.00	0.42
AME055	2	1.54	0.53	0.00	0.35	2	1.26	0.36	0.00	0.21	2	2.00	0.69	0.00	0.50
AME076	2	1.97	0.69	0.00	0.49	2	1.91	0.67	0.00	0.48	2	2.00	0.69	0.00	0.50
AME102	4	3.37	1.29	0.07	0.70	3	1.42	0.57	0.04	0.30	4	3.05	1.22	0.06	0.67
AME154	3	2.60	1.02	0.00	0.61	3	1.83	0.79	0.00	0.45	3	2.27	0.94	0.00	0.56
AME177	6	2.75	1.19	0.74	0.64	5	2.55	1.08	0.72	0.61	6	2.76	1.18	0.73	0.64
AME178	9	5.38	1.89	0.00	0.81	5	2.70	1.14	0.00	0.63	9	4.42	1.72	0.00	0.77
MAMA1	5	2.77	1.29	0.00	0.64	4	2.21	1.03	0.00	0.55	5	2.53	1.22	0.00	0.61
MAMA5	8	4.37	1.71	0.01	0.77	4	2.64	1.09	0.00	0.62	8	3.91	1.64	0.01	0.74
MAMA11	5	2.99	1.23	0.00	0.67	2	1.40	0.46	0.00	0.29	5	2.49	1.07	0.00	0.60
MAMA13	4	3.02	1.22	0.01	0.67	4	2.74	1.17	0.00	0.64	4	3.62	1.32	0.01	0.72
Mean	4.3	2.74	1.09	0.09	0.59	3.3	2.01	0.78	0.10	0.44	4.4	2.62	1.06	0.09	0.58
SD	0.6	0.29	0.11	0.06	0.04	0.3	0.18	0.09	0.07	0.05	0.6	0.23	0.10	0.06	0.03

Na, number of observed alleles; Ne, effective number of alleles; I, Shannon information index. Ho, observed heterozygosity; and He, Nei´s diversity index.

3.8 Genetic diversity

The overall genetic diversity of the 132 oat genotypes was intermediate (He = 0.58 ± 0.03); diversity was greater in Pop 1 (He = 0.59 ± 0.04) than in Pop 2 (He = 0.44 ± 0.05) (Table 4). Extending the analysis per population, the observed alleles per marker ranged between 2 and 9, and between 2 and 5, and the total observed alleles over all markers were 61 and 46, for Pop 1 and Pop 2, respectively. The expected allele number per locus varied from 1.57 to 5.16 in Pop 1, and between 1.05 and 3.26 in Pop 2 (Table 4). A total of 14 private alleles were detected in Pop 1 with the SSRs AME102 (N = 1), MAMA_11 (N = 3), AME178 (N = 4), MAMA_1 (N = 1), MAMA_5 (N = 4), AM22 (N = 1); this resulted in 40 oat genotypes containing one to three private alleles (Supplementary Table S2). Allele frequencies are provided in Supplementary Figure S2.

The observed heterozygosity (Ho) was near zero for the majority of SSRs, with exception of AM14 (Ho = 0.48), and AME177 (Ho = 0.73) (Table 4). The heterozygosity fluctuated between 0 and 0.33 in the 132 individual oat genotypes, with a mean of 0.09, indicating a low degree of genetic segregation and/or contamination, considering that the DNA was obtained from the combined extraction of 10 seedlings per genotype (Supplementary Table S2).

3.9 Phenotypic value and diversity of the germplasm by groups

To obtain an overview of the germplasm in terms of phenotypic performance and diversity, we applied non-parametric comparisons between phenotypic clusters, genetic populations, in categorical groups by geographic origin (Chilean, foreign), and selection for different uses (commercial food-oat breeding and pre-breeding). The phenotypic diversity (H´) was higher in Cluster 2 (H'= 0.62 ± 00.18) and Cluster 3 (H' = 0.60 ± 0.17) than in Cluster 1 (H' = 0.21 ± 0.06), in Pop 1 (H'= 0.63 ± 0.07) than in Pop 2 (H'= 0.57 ± 0.11), in Chilean (H' = 0.64 ± 0.18) than in the foreign (H' = 0.55 ± 0.12) germplasm, and in modern (H' = 0.62 ± 0.15) than in historic (H' = 0.55 ± 0.12) germplasm (Figure 5A). As expected, phenotypic diversity was lower in the food-oat breeding germplasm (H' = 0.42 ± 0.15) than in the pre-breeding (H' = 0.65 ± 0.19) group.

The phenotypic performance of the germplasm for food-oat breeding was better in Cluster 2 (MGIDI = 5.34 ± 0.91) and Cluster 3 (MGIDI = 5.63 ± 1.00), than for Cluster 1 (MGIDI = 7.91 ± 1.13); there was a tendency to a better performance in Cluster 2 than in Cluster 3 ($P = 0.071$) (Figure 5B). Also, the phenotypic performance was superior in Pop 2 (MGIDI = 5.19 ± 0.95) than Pop 1 (MGIDI = 5.97 ± 1.19), in Chilean (MGIDI = 5.34 ± 1.08) than in foreign (MGIDI = 6.06 ± 1.14) germplasm ($P < 0.001$), and in modern (MGIDI = 5.34 ± 6.36) than in historic (MGIDI = 6.36 ± 1.12) germplasm ($P < 0.001$). As expected, phenotypic performance was higher in the food-oat breeding (MGIDI = 4.23 ± 0.39) than in pre-breeding (MGIDI = 6.03 ± 1.00) selected germplasm ($P < 0.001$).

The genetic diversity was similar ($P > 0.05$) in the three phenotypic clusters, between the Chilean (He = 0.53 ± 0.11) and foreign (He = 0.59 ± 0.16) germplasm, and between historic (He = 0.54 ± 0.14) and modern (He = 0.59 ± 0.13) germplasm (Figure 5C). As expected, genetic diversity was higher ($P = 0.035$) in Pop 1 (He = 0.59 ± 0.04) than Pop 2 (He = 0.44 ± 0.05). A tendency to lower

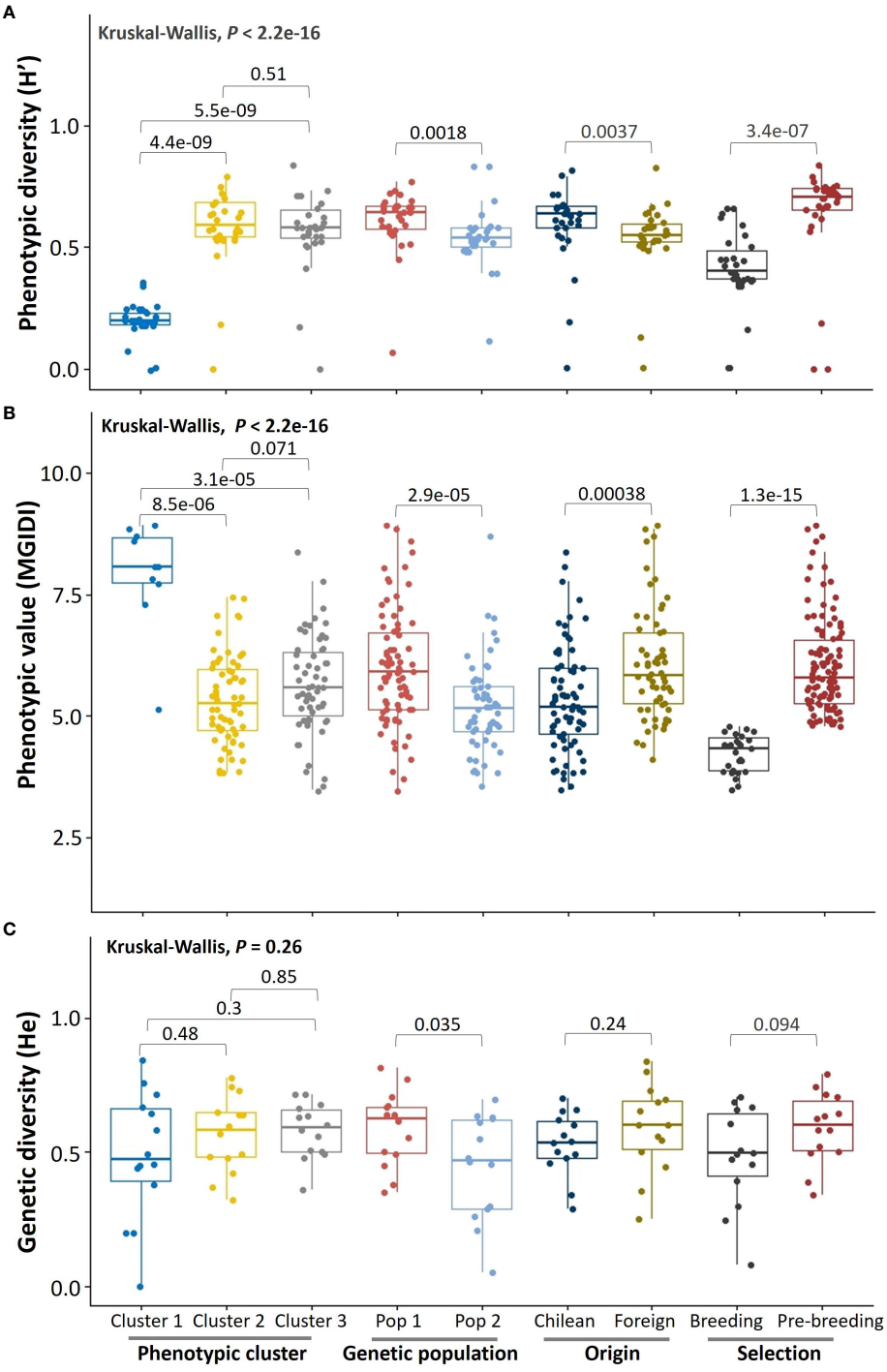


FIGURE 5
Overall phenotypic performance and diversity of the germplasm by groups. **(A)** Shannon-Weaver scaled phenotypic diversity, **(B)** multi-trait genotype-ideotype index, and **(C)** Nei's genetic diversity index. The error bars are the minimum and maximum values; the horizontal line in the box is the median. The Chilean germplasm corresponds to genotypes created in Chile using diverse origin germplasm.

genetic diversity ($P = 0.09$) was observed in the germplasm selected for food-oat breeding (He = 0.48 ± 0.18) compared to the pre-breeding (He = 0.59 ± 0.13) germplasm. Different number of private alleles were found; they were present in Cluster 1 (N = 3) and Cluster 2 (N = 1), Pop 1 (N = 14), foreign (N = 6) and Chilean (N = 2) germplasm, in pre-breeding (N = 17) selected germplasm and in historic (N = 6) genotypes (Supplementary Table S2). However, private alleles were absent in Pop 2, Cluster 3, modern germplasm and in food-oat breeding selected germplasm (Supplementary Table S2).

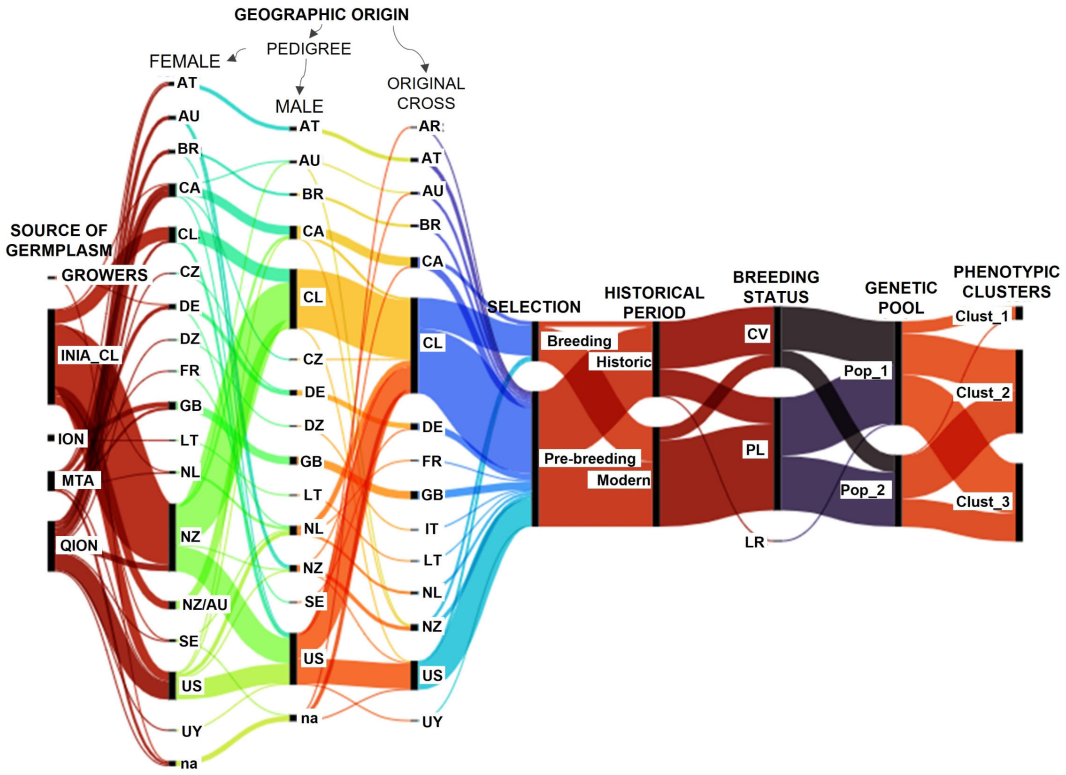


FIGURE 6
Relationships of the oat genotypes according to their geographic origins and grouping patterns. QION, Quaker International Oat Nursery; ION, International Oat Nursery; MTA, material transfer agreement. CV, cultivar; PL, pure line; LR, land race. R, Argentina; AT, Austria; AU, Australia; BR, Brazil; CA, Canada; CL, Chile; CZ, Czech Republic; DE, Germany; DZ, Africa; FR, France; GB, United Kingdom; IT, Italy; LT, Lithuania; na, not assigned; NL, Netherlands; NZ, New Zealand; SE, Sweden; US, United States of America and UY, Uruguay. The Chilean germplasm corresponds to genotypes created in Chile using diverse origin germplasm. Historic: germplasm created on or before 2010, and modern: after 2010.

The alluvial diagram allowed us to visualize the origins of the oat genotypes and group them according to their phenotypic similarity and genetic pools (Figure 6). The germplasm showed a complex interrelated pattern, reflecting a germplasm mostly created from parents with diverse origins. The germplasm was formed mainly by Chilean accessions and others from the USA and Canada, and a low proportion from other countries of South America (Argentina, Uruguay, Brazil), Europe (Germany, France, United Kingdom, Italy, Lithuania, Netherland, Austria), and Oceania (New Zealand, Australia), representing 15 different geographic origins.

Thirty-six out of 46 cultivars, 20 out of 86 pure lines and the Rubia Corriente landrace are historic genotypes created 2010 or earlier (Figure 6). Cultivars were grouped in higher proportion in Pop 1 (N = 33) than Pop 2 (N = 13), while pure lines were in similar proportions in Pop 1 (N = 44) and Pop 2 (N = 41), resulting in a higher proportion of historic germplasm in Pop_1 (59%) than Pop_2 (20%). Pop_1 contained nine out of 10 genotypes from Cluster 1, and a slightly lower number of genotypes from Cluster 2 (N = 31) than Cluster 3 (N = 38), while Pop 2 contained only one genotype from Cluster 1 and more genotypes from Cluster 2 (N = 32) than Cluster 3 (N = 21), explaining the better phenotypic performance of Pop 2 in relation to Pop 1. Consequently, the germplasm selected for food-oat breeding corresponded mainly to modern germplasm belonging to Pop 2 and Cluster 2.

4 Discussion

Oats have gained popularity worldwide since oat-based products consumption can contribute to lower cholesterol and diabetic effects, preventing disease and promoting human health (Paudel et al., 2021). This cereal ranks second in cultivated area in Chile; it is mainly an export product (ODEPA, 2023), and as such is economically important for local oat producers and processing industries. During 2010-2023 period, 70 to 90% of the Chilean oat cropping area has been covered with cv. Supernova INIA (De la Fuente, 2022). The wide spread of dominant cultivars, generating continuous low crop genetic diversity and high uniformity, could lead to crop vulnerability (Khoury et al., 2022; Salgotra and Chauhan, 2023), thus the integration of new genetic diversity associated with favorable commercial traits is essential to increase resilience against climate change (Swarup et al., 2021), and to deal with the emergence of new biotic and abiotic stresses (Ristaino et al., 2021; Skendžić et al., 2021; FAO, 2022). A way to increase diversity

in modern cereal cultivars, highly uniform and sown in large areas like oats, is the quick turnover of cultivars (Khoury et al., 2022). However, none of the cultivars released in Chile and new advanced lines created have significantly outperformed Supernova INIA, resulting in low adoption by producers and a low release rate of new cultivars. Therefore, quantifying and broadening the available genetic diversity are expected to be critical to increase the rate of cultivar turnover in Chile.

A study of genetic diversity of the oat germplasm available in INIA, based on molecular markers and including new advanced lines and commercial cultivars was missing, because the selection was mostly based on agronomic performance. A comparable phenotypic evaluation of the germplasm, including historic and modern genotypes of diverse origin, was also missing due to disconnected evaluations in different historical periods. For these reasons, we analyzed the phenotypic and genetic diversity of 132 oats accessions with 28 traits and 14 SSR markers, in a field trial sown in La Araucanía Region, where the main oat cropping area is concentrated. Important genetic parameters affecting the success of genetic breeding were also estimated. The results obtained here supported the applicability of most of the current food-oat breeding objectives, and consequently promise good prospects for obtaining new cultivars in the short term. However, some noteworthy issues were identified which could make breeding from the currently available germplasm difficult in the long term.

Moderate genetic diversity and a discrete population structure were found in the studied germplasm, similar to most foreign germplasm analyzed worldwide (Nersting et al., 2006; Jan et al., 2020; Lyubimova et al., 2020; Wang et al., 2023), showing a low rate of diversification of the germplasm. The Chilean germplasm showed similar genetic diversity in comparison to foreign germplasm from 14 different origins. The phenotypic diversity of the germplasm in key agronomic and grain quality traits accounted for a low proportion of genotypes with favorable phenotype in relation to the reference cultivar Supernova INIA, showing a high grade of fixation of these traits under intensive selection in the breeding program. These factors could explain the relatively low success of the breeding efforts taking place in Chile. Despite these negative aspects, a group of 26 superior genotypes compared to Supernova INIA, mostly modern Chilean pure lines, were selected by a multi-trait index calibrated for commercial food-oat breeding. These genotypes could play an important role in widening the genetic diversity of the oat crop in Chile if some of them pass the necessary tests for release as new cultivars.

The rest of the genotypes of lower performance for food-oat breeding compared to Supernova INIA are a valuable genetic resource for pre-breeding and further studies on other characters not addressed in the present study. Thus, the conservation of this germplasm is relevant. These results emphasize the key role of using diverse origin germplasm and carrying out genetic breeding in Chile, since almost all the selected superior genotypes were created with foreign germplasm but selected in the southern Chile environment. It has been proposed that investment in public breeding programs, providing pre-breeding and other diversification services to formal seed systems, also of advanced breeding technologies (e.g. genetic modification and gene editing), will be needed to mitigate negative impacts on modern cultivar diversity (Khoury et al., 2022).

4.1 Phenotypic variation and heritability of the traits

Phenotypic variation is the distribution or range of morphological, phenological, developmental and biochemical traits that are expressed in individual taxa (Kalisz and Kramer, 2008), determining the capacity of plants to adapt to changing environments and to colonize new habitats (Boquete et al., 2021). Having populations with a high range of phenotypic variability is indispensable to achieve genetic gain in breeding programs (Swarup et al., 2021). For this reason, we evaluated the phenotypic variation of the germplasm in terms of dispersion of the data (genetic coefficient of variation, CVg) and range of variation. Low CVg (< 6%) was found in hectoliter weight, groat content and heading days, while the 25 other traits exhibited moderate to high CVg, in most cases in an adequate range of variation for a selection program. The range of variation in phenotypic traits can broadly vary depending on the germplasm and environmental conditions. For example, the ranges of variation observed in heading days (79.33-97 d, 41-69 d), thousand hulled grain weight (6-27.33 g, 19-45.7 g) and plant height (52-134.7 cm) in 38 oat Indian genotypes (Kumar et al., 2023), and in 64 European oat cultivars and 17 landraces (Nersting et al., 2006), respectively, were of lower magnitude in comparison to the present study.

Heritability is a key concept in breeding, corresponding to the fraction of the total variance among plants of a population that may be attributed to genetic differences between them, which is one of the most important components of the breeder's equation that aims to predict the expected response to selection (Mazurkievicz et al., 2019). Most traits studied here exhibited moderate to high heritability (H^2 > 0.50), indicating a good prospect to achieve genetic gain in a selection program. The larger the heritability of a trait, the greater the expected genetic gain, in which case artificial selection can be carried out more efficiently (Mazurkievicz et al., 2019). The low heritability (H^2< 0.40) observed for grain yield, plant height, vigor scores at tillering, the incidence of diseases like *P. syringae*, Barley Yellow Dwarf Virus, and hull staining, reflects the large influence of environmental factors affecting these traits. Selection may be difficult in traits with low heritability (< 0.40), due to the confounding effect of the environment (Majhi, 2020).

Several studies on heritability of oat traits have been published, with heading days, plant height at maturity, and thousand hulled grain weight, being the most commonly traits investigated (Sürek and Valentine, 1996; Yan et al., 2016; Kumari et al., 2017; Premkumar et al., 2017; Chauhan and Singh, 2019; Leišová-Svobodová et al., 2019; Mazurkievicz et al., 2019; Meira et al., 2019; Haikka et al., 2020; Vanjare et al., 2021; Brzozowski et al., 2022). However, heritability cannot be generalized to a crop since it is highly specific, and valid only for the material involved in the experiment and the experimental environment (Majhi, 2020). High variation in the results has been observed in oats, depending on the

type of germplasm, management, and experimental design of the trials (Supplementary Table S9). The heritability in other studies included 12 of the 28 traits studied here, including heading days (0.46-0.94), plant height at maturity (0.50-0.98), panicle length (0.12-0.86), thousand hulled (0.26-0.97) and thousand dehulled (0.30-0.94) grain weights, grain yield (0.00-0.88), groat protein (0.88-0.93), groat fat (0.13-0.99), groat content (0.23-0.91), hectolitre weight (0.89), panicle type (0.68) and hull color (0.88); the values obtained here were mostly within the ranges observed in aforementioned studies (Supplementary Table S9). We are contributing with broad sense heritability data for 16 oat traits not reported in the reviewed literature.

4.2 Phenotypic diversity

We used the scaled Shannon-Weaver diversity index (H') to quantify the phenotypic diversity of the germplasm on mostly quantitative traits, except for panicle type and hull color, which are qualitative traits. A low H' indicates extremely unbalanced frequency classes for an individual trait and a lack of diversity (Upadhyaya et al., 2002). Almost all the traits exhibited high phenotypic diversity (H' > 0.6), with exception of hectoliter weight and groat dry matter (H' < 0.3). Hectoliter weight, groat content and thousand hulled grain weight, which are important quality traits for the industry, are mostly fixed in the germplasm in relation to the reference cultivar Supernova INIA. Consequently, these results show possibilities of maintaining but not substantially increasing the quality of future cultivars.

There was high diversity in heading days (H' > 0.7) but the germplasm was mainly in the range of midseason to long cycles, which might make it hard create earlier cultivars. Grain yield and plant height at maturity showed a high diversity; about 15% of the germplasm had higher yield and shorter plant height compared to Supernova INIA. However, since grain yield exhibited low heritability, low genetic gain is expected. The disease scores and new implemented grain quality traits such as high severity and low severity groat staining, broken grains after peeling and groat protein and fat contents had high diversity, with a good proportion of genotypes with favorable phenotype *versus* Supernova INIA, thus there are good prospects for improvement of these traits in new cultivars.

There are few studies in oats reporting diversity indexes. For instance, diversity (H) was lower (0.9-1.4) in a group of 84 Nordic oat cultivars and landraces, considering three traits (Nersting et al., 2006). In a group of 91 Polish oat landraces which included eight traits, H was only slightly lower (0.79-2.08) (Boczkowska et al., 2016) than in this study (0-2.55). The studied oat germplasm exhibited high overall phenotypic diversity (H' > 0.6) (Yemataw et al., 2018), considering a wider number of phenotypic traits and oat genotypes compared to the available literature.

4.3 Pairwise correlations between traits

Quantitative traits often exhibit complex inter-relationships which can hinder breeding for multiple traits at a time. For this reason, understanding the associations between traits is a prerequisite to approach breeding (Yan et al., 2016). Non-zero genetic correlations can occur by close linkage of loci, or by pleiotropy which occurs when two traits are controlled by the same loci (Bernardo, 2020). Unlike linkage-correlations, pleiotropic-correlations cannot be dissipated by repeated cycles of meiosis, due to their physiological bases (Bernardo, 2020). The ability to differentiate between pleiotropy and close-linkage correlations will determine the optimum breeding strategy (Chen and Lübberstedt, 2010).

The nature of the correlations in oats is still poorly understood, shown by the scarcity of related literature. Thus, with the purpose of generating validated information on diverse origin oats, the pairwise Pearson correlations (r) between all the traits routinely measured in the oat breeding program were calculated, resulting in 108 weak to moderate significant associations ($P \leq 0.05$). However, a significant value of r does not imply causality or a direct relationship between the variables, since two variables can be correlated because both have a significant degree of correlation with a third variable (Armstrong, 2019). To avoid this confusion, partial correlation coefficients were computed to corroborate the results. Partial correlations have been useful when multiple variables are present because they consider the effect of a third confounding variable on the correlation (Armstrong, 2019; Olivoto and Lúcio, 2020). As expected, only 44 significant associations were confirmed by the partial correlation analysis, the remaining 64 associations were caused by other confounding variables. Interestingly, 18 significant partial correlations were not detected by pairwise correlations, showing the ability of the partial analysis to detect associations hidden because of other confounding variables.

The trait correlations found in most studies used fewer traits and different comparisons (Khan et al., 2014; Crestani et al., 2015; Boczkowska et al., 2016; Baye et al., 2020), making comparison to our results difficult. The significant associations we found with pairwise and partial correlations were mostly favorable for breeding purposes, indicating good possibilities to improve these traits simultaneously, with exception of groat protein content, which showed unfavorable associations with key traits like grain yield. Similar results for the negative association between grain yield and groat protein (r = -0.38, $P < 0.01$) and opposite result for the negative association between protein and groat content (r = 0.13, $P < 0.05$), were observed in two biparental oat populations in different environments and seasons; there were descendants with appropriate combinations of these traits and transgressive segregation (Yan et al., 2016). Since the unfavorable associations here were of small biological importance, they should not be a limitation for genetic improvement if good combinations of traits are identified in the germplasm, together with a continued effort to develop the desired cultivars (Yan et al., 2016).

4.4 Phenotypic clustering of the oat germplasm

A principal component analysis showed the distribution of the genotypes (distances in the factor map) and the combination of the

traits in the oat germplasm (correlations). The high phenotypic diversity of the germplasm was evident. A cluster analysis of the principal coordinates grouped the genotypes exhibiting similar phenotypic patterns into three clusters. Depending on whether the trait mean in the cluster was higher or lower compared to the overall mean of the total population, the trait was considered favorable or unfavorable for food-oat breeding. Cluster 1 grouped mainly oats with aptitude for feed, markedly different than the rest of genotypes due their very high plant heigh and straw weakness, such as the landrace Rubia Corriente widely used in Chile as fodder, silage, and grain in animal production (Beratto, 1977), among others.

The remaining germplasm was separated in similar proportion in Cluster 2 and Cluster 3. Cluster 2 had high values of grain yield, agronomic scores, industrial grain quality and longer open panicles, whereas Cluster 3 grouped genotypes with low plant height and low lodging, high protein content, low industrial grain quality, and short compact panicles. These two phenotypic patterns show limitations to breed dwarf and lodging-tolerant varieties combined with high yield and high industrial grain quality, since these phenotypes were not observed in the germplasm, independently of the breeding status (pure lines, cultivars), geographic origins (Chilean, foreign), and historical period (historic, modern). Creating a fourth group, combining positive characteristics of Cluster 2 and Cluster 3, would allow diversifying the phenotypes of the available oat germplasm.

4.5 Phenotypic value of the oat germplasm for food-oat breeding

We estimated the BLUP phenotypic values of the germplasm, corresponding to the best linear neutral prediction for estimating genetic competence (Mahdi and Mohammad, 2022), accounting for both the additive and nonadditive genetic effects of a line (Henderson, 1976). This is critical information for parental selection decisions and for determining the relative "eliteness" of a line (Cobb et al., 2019). The simultaneous selection for multiple traits and different breeding objectives made it hard to visualize the eliteness of the oat genotypes. For this reason, we used the MGIDI, a BLUP-based selection index which carried the correlations between traits to a single plane, considering the breeding objectives (increase-decrease), and the intra-mean traits heritability in the estimation of genetic gain (Olivoto and Nardino, 2021).

Only 26 genotypes, 20% of the total germplasm, outperformed the elite phenotype of the reference cultivar Supernova INIA for food-oat breeding. These genotypes would be valuable in the diversification of oat crops in Chile if some of them are released as cultivars and/or used in crosses. However, the lower diversity of the selected genotypes compared to the rest of the germplasm, and the qualities for breeding in the selected group in relation to Supernova INIA, confirmed the convenience of a more diverse germplasm in favorable traits than that currently available in the breeding program at INIA-Chile.

The germplasm with lower performance than Supernova INIA was discarded for commercial breeding but selected for pre-breeding, because these genotypes exhibited specific traits useful for development of new cultivars. The introduction of genetic diversity from genotypes that have contrasting phenotypic traits is a major challenge; there are numerous examples in plants showing an unfavorable phenotype due to adverse genetic background effects and linkage drag with the desired trait (Swarup et al., 2021). Since the INIA breeding program is mainly based on bi-parental crosses, the descendants will contain up to 50% of the genome of each parent, unfavorable phenotypes being expected using parents of low phenotypic performance. Consequently, a long-term pre-breeding process will be required to reduce the negative effects in new cultivars.

4.6 Population structure

The clustering of the oat germplasm based on genetic distance estimated with 14 SSR markers separated the germplasm into two genetic subtrees, although with low bootstrap support. A Bayesian approach that estimates for each accession the proportion of the genome that originates from each subpopulation, also called percentage of admixture (Pritchard et al., 2000; Montilla-Bascón et al., 2013), together with Evanno´s statistic (Evanno et al., 2005), detected the existence of two different subpopulations or genetic pools, called Pop 1 and Pop 2; Pop 1 included Rubia Corriente, Llaofén INIA, Eva and Júpiter INIA, while Pop 2 was the genetic pool of Supernova INIA and Urano INIA. Interestingly, Pop 1 grouped most historical and foreign germplasm, whereas Pop 2 was mostly composed of modern Chilean pure lines, commercial cultivars available in Chile and a few foreign cultivars, suggesting that allele combinations in Pop 2 would be associated with better agronomic performance in the southern Chile environment.

The oat germplasm exhibited a weak population structure, with an optimal number of two sub-populations. A Structure analysis on a collection of 141 *Avena sativa* L. landraces including 110 white, and 31 red oats from Spain based on 31 SSRs supported the existence of two gene pools (Montilla-Bascón et al., 2013). The same happened with 24 landraces from India studied with 24 SSRs (Rana et al., 2019); in 85 oats with white, yellow, and brown seeds as well as a subgroup of naked oats from 18 different countries based on seven SSRs (Havrlentová et al., 2021); in 91 indigenous accessions from Poland using eight ISSR markers (Boczkowska et al., 2016); on 288 oats genotypes of diverse origin using 2143 SNPs (Wang et al., 2023); in 487 *Avena sativa* accessions mainly from Poland using 7411 SNPs (Koroluk et al., 2023); and in 38 oat accessions from India using 22 ISSRs (Kumar et al., 2023). Structure supported three subpopulations in a group of 1,000 world-wide oat accessions, including cultivars, germplasm of uncertain improvement status, and landraces, based on data from 2,715 SNP markers (Winkler et al., 2016), and in a 260 diverse origin collection of husked, naked and black oats using 15 SSRs (Leišová-Svobodová et al., 2019). Thus, in different oat germplasms and applying different types and number of markers, the population structure was essentially formed by two genetic pools. The low level of differentiation between populations shows a low level of diversification of the oat germplasm. This might be due to the recent domestication of *Avena sativa* L., which appeared in

cultivation several thousand years later than wheat and barley (Zohary and Hopf, 2000).

4.7 Genetic variation and diversity

Genetic variation is related to differences in particular DNA sequences between individuals, while genetic diversity is related to DNA differences in populations (Swarup et al., 2021). It is said that permanent access to genetic variation for different phenotypic traits is a requisite for obtaining long-term breeding progress (Swarup et al., 2021), whereas genetic diversity is the main driving force for the selection and evolution of populations (Salgotra and Chauhan, 2023). The oat genotypes were well differentiated, showing variability based on their genetic distances except for two non-differentiable duplicates. The average PIC (0.58), average number of alleles per locus (4.3) and number of rare alleles per locus (1.42), were lower than in other studies, such as 177 white and red oats land races and cultivars from Spain characterized with 31 SSRs, which had an average PIC of 0.80, 14.45 alleles and 4.45 rare alleles per locus (Montilla-Bascón et al., 2013). The average genetic diversity of the germplasm was moderate (He = 0.52 ± 0.03); it was 15% to 28% higher in Pop 1 than in Pop 2, depending on the index. Pop_1 retained 100% of the allele richness *versus* Pop 2 accounting for 75.4%; private alleles were only found in Pop 1. Thus Pop 1 is a reservoir of genetic diversity for oat pre-breeding.

The genetic diversity of this oat germplasm was difficult to compare with other studies due to the different molecular markers used, population sizes and origins. Having said that, similar genetic diversity (He) compared to the present study was obtained in a group of 16 exotic oat genotypes from Europe and Pakistan (0.12-0.53) using five RAPD primers with 23 loci amplified (Jan et al., 2020); in 18 oat cultivars from Russia, Norway, Netherlands, and Sweden (0.33-0.75) based on avenin-like alleles (Lyubimova et al., 2020); in 64 oat cultivars from Europe (0.5-0.63) using seven SSRs (Nersting et al., 2006); in 288 oat genotypes from diverse origin (0.096-0.50) using 2143 SNPs (Wang et al., 2023); and in a 260 naked, husked and black oats collection (0.48-0.61) using 15 SSRs (Leišová-Svobodová et al., 2019). However, low genetic diversity was found in 23 oat cultivars from Poland (0.20) using the dominant markers ISSR, RAPDs and AFLPs (Boczkowska et al., 2016), in 177 red and white oats from southern Spain (0.29) using 31 SSRs (Montilla-Bascón et al., 2013), in 72 Polish oats (0.15-0.30) using 36 ISSRs (Koroluk et al., 2022), and in 60 accessions representing 13 *Avena* species (0.000-0.068) using retrotransposon primer binding sites-iPBS (Androsiuk et al., 2023). Thus, the mostly moderate genetic diversity observed here, which is coincident with most other studies from around the world, shows a limitation for oat breeding, indicating the need to increase the diversity using different strategies to ensure genetic gain in the long term.

Most studies of genetic diversity in oats, including the present work, used binary data to estimate genetic parameters and population structure, applying different types of markers (SSRs, ISSR, RAPD, AFLP, iPBS), masking the size and sequence of alleles. In the case of SSRs, the length of the alleles has served to approximate the number of repetition units, and has been used to calculate genetic and evolutionary distance between individuals (Šarhanová et al., 2018). Additionally, SSRs fragments of the same size but with different sequences, frequently referred to as size homoplasy, have been revealed through sequencing (Estoup et al., 2002). Thus, the use of binary data could cause alleles misinterpretation and bias in genetic parameters. For example, in 1135 samples of different populations of *Ceratonia siliqua* (Leguminosae), sequence-based SSRs genotyping allowed for a better estimate of population divergence, detecting higher private allele richness compared to size fragments scoring (Viruel et al., 2018). In another study, SSRs fragment sequences revealed higher number of alleles and higher genetic diversity, but a similar F_{ST}, in comparison to fragment size scoring in 384 accessions of *Donatia fascicularis* (Stylidiaceae), 88 accessions of *Mulguraea tridens* (Verbenaceae), and 384 accessions of *Oreobolus obtusangulus* (Cyperaceae) (Šarhanová et al., 2018). On the other hand, a thorough review of homoplasy in diverse molecular ecology studies, concluded that homoplasy does not represent a significant problem in most types of analyses of population genetics, because it is often compensated by a large amount of variability in microsatellite loci (Estoup et al., 2002). Therefore, whilst including information of alleles sequences or size might have allowed detecting greater variation and improving the accurateness of populations structure and genetic diversity estimates, we consider the current results as reliable. For the purposes of the present work, the SSRs markers coded as binary data were a cost-effective way to estimate genetic parameters in oats, accepting that a rate of unknown bias might exist. We are not aware of any published studies quantifying homoplasy in oats. Comparing different markers platforms, SSRs still represent a useful marker system because of their high mutation rates and cost-effectiveness (Viruel et al., 2018), being recommended for projects with limited budgets (Jennings et al., 2011), such as the present one.

4.8 Perspectives for oat breeding in Chile

Oat breeding in Chile has had slow progress after the release of cv. Supernova INIA, although the breeding program has been permanently introducing new germplasm of diverse origin. The similar genetic diversity, and higher phenotypic diversity and value of the Chilean germplasm, compared to the available foreign germplasm studied here, could explain the relatively poor success. Similar diversity status and higher phenotypic performance was observed in modern germplasm compared to historical genotypes. This reflects the efforts made in the last decade by the INIA breeding program aiming to release new lines with a better performance in comparison to imported and historic cultivars. Maintaining genetic diversity in a breeding program is essential to guarantee sustainable genetic gain for targeted traits (Salgotra and Chauhan, 2023; Sanchez et al., 2023). Both the Chilean and modern germplasm studied here conserved similar genetic diversity in comparison to the foreign and historical germplasm, harboring a similar magnitude of diversity in relation to other oat germplasms evaluated in other studies. Thus, a good prospect for genetic

improvement is seen in the short term if the use of these genetic resources is optimized. The introduction of new sources of diversity is relevant in this regard, to avoid reaching a genetic gain plateau in the longer term (Sanchez et al., 2023).

The germplasm studied here exhibited two genetic subpopulations with a weak differentiation, showing a low genetic divergence in the germplasm, and reflecting the short domestication period of oats (Zohary and Hopf, 2000). Genetic divergence can be due to mutation, genetic drift, and selection (Kozak et al., 2011), and is important in plant breeding to identify and utilize genetic variation within and between populations to develop new and improved varieties (Dhanalakshmi et al., 2023). Pop 1 was identified as a reservoir of allele richness and genetic diversity, whereas Pop 2 showed slightly lower genetic diversity, probably caused by elimination of inferior alleles during selection, deduced from their superior overall phenotypic performance compared to Pop 1, and the lack of private alleles in Pop 2.

If the divergence between subpopulations is neutral regarding the frequency of favorable alleles over multiple loci, the best hybrids are more likely to come from inter-population crosses (Mackay et al., 2021). Despite the higher phenotypic performance of Pop 2, examples of oats with good phenotypic value were also found in Pop 1, which would be good parental candidates for implementing between-population crosses, and at the same time achieve a higher dispersion of favorable alleles. The dispersion of favorable alleles between parents causes transgressive segregation, in which progenies exhibit a phenotype outside the range of the parents (Mackay et al., 2021). Importantly, phenotypes produced by transgressive segregation are heritably stable and can be observed in crosses involving parents of proximal phenotypes, as for example heading days in rice, due to the existence of hidden genetic variation between the parents (Koide et al., 2019). Several foreign pure lines included in this study exhibited acceptable phenotypic values and belonged to different genetic subpopulations. These lines have not yet been used in crosses, or in other cases, their progenies are still in preliminary evaluation stages in Chile. The selection of transgressive genotypes in crosses implemented with the information generated in this study, should be the next step for generating new improved lines in the short term with the available germplasm. It is projected that this strategy would be useful to express transgressivenness in important characters for food-oat breeding, exhibiting a similar but not superior phenotype compared to supernova INIA in the studied germplasm, such as hectoliter weight, heading days, groat content, and plant height, among others.

A two-way strategy is proposed for long term breeding, based on the observed results. First, the enrichment of the genetic diversity to overcome the diversity plateau can be approached through a continuous exchange of germplasm, which would be dependent on the accessibility of new genetic resources. This path can also be explored through the creation of new genetic diversity through strategies such as genetic modification, and gene-editing and mutation-breeding. Transformation of oats using biolistic bombardment improved the tolerance to osmotic stress in transgenic plants (Maqbool et al., 2009). Gene-editing has been successfully applied in the breeding of wheat, rice, and barley to improve tolerance to biotic and abiotic stresses, grain quality and yield (Riaz et al., 2022). However, despite the usefulness of these modern techniques in genetic breeding, they are not currently accepted for food production by regulatory agencies in Chile. Alternatively, mutation is a non-transgenic powerful approach to generate novel genetic variation that can be exploited by breeding programs (Abaza et al., 2020). At least two thousand rice, barley, wheat, soybean, maize, and oat mutant varieties have been released to farmers (https://nucleus.iaea.org/sites/mvd/SitePages/Home.aspx). The development of Targeting Induced Local Lesion IN Genomes (TILLING), consisting in mutagenesis followed by a rapid identification of mutations in the genes of interest (Szurman-Zubrzycka et al., 2023), has been broadly used to improve traits such as starch synthesis, plant architecture, disease resistance, drought and salinity tolerance, and other yield parameters in cereals (Irshad et al., 2020; Nouman-Khalid et al., 2021; Abdelnour-Esquivel et al., 2010). In oats, TILLING allowed to induce variation in genes encoding for enzymes involved in the pathways of lignin and beta-glucan biosynthesis (Chawade et al., 2010). Recently, mutagenesis by direct current electrophoresis bath (DCEB) was investigated in rice (Zou et al., 2023). Mutation by exposing seeds or plants to cosmic radiation in outer space, or space-breeding, has allowed the release of at least 66 varieties in China (Liu et al., 2009; Mohanta et al., 2021). Second, the breeding program can be optimized through continuous monitoring of existing genetic diversity, improving selection of parents in crosses to maintain a high diversity level. It is well known that investment in oat breeding programs on a global scale has been rather low, as shown by fewer published articles related to new breeding technologies in the species compared to other cereal crops like wheat and barley. This fact might limit the implementation of modern breeding tools in oats also in Chile.

5 Conclusion

The oat germplasm studied here contained high phenotypic diversity but with a discrete proportion of genotypes exhibiting adequate phenotypic performance for food-oat breeding compared to Supernova INIA, the most cropped cultivar in Chile. This, together with the higher phenotypic performance and similar genetic diversity of the Chilean germplasm compared to foreign germplasm, explain in part the slow progress of Chilean breeding after Supernova INIA, even with continuous introduction of new germplasm. Heritability, range of variation and correlations of phenotypic traits in the studied germplasm shows an auspicious genetic breeding, for most food-oat breeding objectives. However, the germplasm studied here showed moderate genetic diversity, with two weakly differenced genetic pools, similar to other oat germplasms studied around the world, reflecting the low genetic divergence in the species. These factors underline the urgent need to enrich the genetic and phenotypic diversity of the currently available germplasm, making efficient use of the genetic resources, integrating the results obtained here in making decisions to maintain high diversity in the breeding program. In the long

term, investing in modern breeding tools or mutation breeding to overcome the diversity plateau in the species, is proposed as an alternative independent of the availability of foreign genetic resources to enrich the species diversity in breeding programs. The results of the present study depict a challenging prospective for oat breeding in Chile.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MM: Formal analysis, Funding acquisition, Project administration, Supervision, Writing – original draft, Conceptualization, Data curation, Investigation. VP: Investigation, Writing – review & editing. MM: Writing – review & editing. FF: Investigation, Writing – review & editing. AV: Investigation, Writing – review & editing. IL: Investigation, Methodology, Writing – review & editing. MS: Methodology, Investigation, Writing – review & editing. RS: Investigation, Writing – review & editing. PH: Investigation, 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 financially by Proyecto Núcleos de Investigación INIA, Instituto de Investigaciones Agropecuarias, Ministerio de Agricultura, Chile.

Acknowledgments

The authors would like to thank to Dr. Stephen Harrison and collaborators of the International Oat Nursery (ION) and the Quaker International Oat Nursery (QION), AAFC Canada, SENOVA Limited, and KWS LOCHOW GMBH, for providing seeds of oat genotypes, and Dr. Lafayette Eaton for English edition and comments to 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.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1298591/full#supplementary-material

References

Abaza, G., Awaad, H. A., Attia, Z. M., Abdel-lateif, K. S., Gomaa, M. A., Safy, M., et al. (2020). Inducing potential mutants in bread wheat using different doses of certain physical and chemical mutagens. *Plant Breed. Biotech.* 8, 252–264. doi: 10.9787/PBB.2020.8.3.252

Abdelnour-Esquivel, A., Perez, J., Rojas, M., Vargas, W., and Gatica-Arias, A. (2010). Use of gamma radiation to induce mutations in rice (*Oryza sativa* L.) and the selection of lines with tolerance to salinity and drought. *In Vitro Cell.Dev.Biol.-Plant* 56, 88–97. doi: 10.1007/s11627-019-10015-5

Achleitner, A., Tinker, N. A., Zechner, E., and Buerstmayr, H. (2008). Genetic diversity among oat varieties of worldwide origin and associations of AFLP markers with quantitative traits. *Theor. Appl. Genet.* 117, 1041–1053. doi: 10.1007/s00122-008-0843-y

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). iMEC: online marker efficiency calculator. *Appl. Plant Sci.* 6, e01159. doi: 10.1002/aps3.1159

Androsiuk, P., Milarska, S. E., Dulska, J., Kellmann–Sopyła, W., Szablińska–Piernik, J., and Lahuta, L. B. (2023). The comparison of polymorphism among *Avena* species revealed by retrotransposon–based DNA markers and soluble carbohydrates in seeds. *J. Appl. Genet.* 64 (2), 247–264. doi: 10.1111/opo.12636

Armstrong, R. A. (2019). Should pearson's correlation coefficient be avoided? *Ophthalmic Physiol. Opt.* 39, 316–327. doi: 10.1111/opo.12636

Arora, A., Sood, V., Chaudhary, H., Banyal, D., Kumar, S., Devi, R., et al. (2021). Genetic diversity analysis of oat (*Avena sativa* L.) germplasm revealed by agro-morphological and SSR markers. *Range Manage. Agrofor.* 42, 38–48.

Baye, A., Berihun, B., Bantayehu, M., and Derebe, B. (2020). Genotypic and phenotypic correlation and path coefficient analysis for yield and yield-related traits in advanced bread wheat (*Triticum aestivum* L.) lines. *Cogent Food Agric.* 6, 1752603. doi: 10.1080/23311932.2020.1752603

Beratto, E. (1977). Efectividad de la selección por línea pura en el mejoramiento de avena Rubia corriente. *Agric. Téc.* 37, 150–155. Available at: https://hdl.handle.net/20.500.14001/36140 (Accessed December 12, 2023).

Beratto, E. (2006). *Cultivo de la avena en Chile* (Temuco, Chile: Colección Libros INIA - Instituto de Investigaciones Agropecuarias. no. 19).

Bernardo, R. N. (2020). *Breeding for quantitative traits in plants. 3rd ed.* (Woodbury, Minnesota: Stemma Press).

Boczkowska, M., Łapiński, B., Kordulasińska, I., Dostatny, D. F., and Czembor, J. H. (2016). Promoting the use of common oat genetic resources through diversity analysis and core collection construction. *PLoS One* 11, e0167855. doi: 10.1371/journal.pone.0167855

Boquete, M. T., Muyle, A., and Alonso, C. (2021). Plant epigenetics: phenotypic and functional diversity beyond the DNA sequence. *Am. J. Bot.* 108, 553–558. doi: 10.1002/ajb2.1645

Brzozowski, L. J., Hu, H., Campbell, M. T., Broeckling, C. D., Caffe, M., Gutiérrez, L., et al. (2022). Selection for seed size has uneven effects on specialized metabolite abundance in oat (*Avena sativa* L.). *G3 Genes Genomes Genet.* 12, jkab419. doi: 10.1093/g3journal/jkab419

Chauhan, C., and Singh, S. (2019). Genetic variability, heritability and genetic advance studies in oat (*Avena sativa* L.). *Int. J. Chem. Stud.* 7, 992–994.

Chawade, A., Sikora, P., Bräutigam, M., Larsson, M., Vivekanand, V., Nakash, M. A., et al. (2010). Development and characterization of an oat TILLING-population and identification of mutations in lignin and beta-glucan biosynthesis genes. *BMC Plant Biol.* 10, 86. doi: 10.1186/1471-2229-10-86

Chen, Y., and Lübberstedt, T. (2010). Molecular basis of trait correlations. *Trends Plant Sci.* 15 (8), 454–461. doi: 10.1016/j.tplants.2010.05.004

Cieplak, M., Okoń, S., and Werwińska, K. (2021). Genetic similarity of *Avena sativa* L. varieties as an example of a narrow genetic pool of contemporary cereal species. *Plants* 10, 1424. doi: 10.3390/plants10071424

Cobb, J. N., Juma, R. U., Biswas, P. S., Arbelaez, J. D., Rutkoski, J., Atlin, G., et al. (2019). Enhancing the rate of genetic gain in public-sector plant breeding programs: lessons from the breeder's equation. *Theor. Appl. Genet.* 132, 627–645. doi: 10.1007/s00122-019-03317-0

Crestani, M., Gonzalez da Silva, JoséA., Woyann, L. G., Zimmer, C., Groli, E., Costa de Oliveira, A., et al. (2015). Correlations among industrial traits in oat cultivars grown in different locations of Brazil. *Aust. J. Crop Sci.* 9, 1182–1189.

De la Fuente, M. C. (2022). "Evaluación de impacto de la variedad Supernova-INIA en la Región de la Araucanía," in *Boletín INIA N° 464, Serie Evaluación de Impacto N° 5* (Santiago, Chile: Instituto de Investigaciones Agropecuarias), 39. p. Available at: https://biblioteca.inia.cl/bitstream/handle/20.500.14001/68616/NR42906.pdf?sequence=7&isAllowed=y (Accessed December 12, 2023).

Dhanalakshmi, T. N., Santosh, D. T., and Shashidhara, N. (2023). "Genetic divergence in plant breeding: forces, markers, and importance for crop Improvement," in *Recent Advances in Agricultural Sciences and Technology*. Eds. N. Biradar, R. A. Shan and A. Ahmad (New Delhi, India: Dilpreet Publishing House), 86–92.

Estoup, A., Jarne, P., and Cornuet, J. M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* 11, 1591–1604. doi: 10.1111/j.1365-294X.2005.02553.x

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

FAO (2022). *FAO Strategy on Climate Change 2022–2031* (Rome: Food and Agriculture Organization of the United Nations), 52. p.

FDA (2023) *Food labeling and nutrition: authorized health claims that meet the significant scientific agreement (SSA) standard, U.S. Food and Drug Administration*. Available at: https://www.fda.gov/food/food-labeling-nutrition/authorized-health-claims-meet-significant-scientific-agreement-ssa-standard (Accessed July 19, 2023).

Fulton, T. M., Chunwongse, J., and Tanksley, S. D. (1995). Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol. Biol. Rep.* 13, 207–209. doi: 10.1007/BF02670897

Haikka, H., Manninen, O., Hautsalo, J., Pietilä, L., Jalli, M., and Veteläinen, M. (2020). Genome-wide association study and genomic prediction for *Fusarium graminearum* resistance traits in Nordic oat (*Avena sativa* L.). *Agronomy* 10, 174. doi: 10.3390/agronomy10020174

Havrlentová, M., Ondreičková, K., Hozlár, P., Gregusová, V., Mihálik, D., and Kraic, J. (2021). Formation of potential heterotic groups of oat using variation at microsatellite loci. *Plants* 10, 2462. doi: 10.3390/plants10112462

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32 (1), 69–83. doi: 10.2307/2529339

Informe de expertos (2023) *Mercado Latinoamericano de Avena, markert report historical and forecasts market analysis*. Available at: https://www.informesdeexpertos.com/informes/mercado-latinoamericano-de-avena (Accessed July 19, 2023).

Irshad, A., Guo, H., Zhang, S., and Liu, L. (2020). TILLING in cereal crops for allele expansion and mutation detection by using modern sequencing technologies. *Agronomy* 10, 405. doi: 10.3390/agronomy10030405

Jan, S. F., Khan, M. R., Iqbal, A., Khan, F. U., and Ali, S. (2020). Genetic diversity in exotic oat germplasm and resistance against barley yellow dwarf virus. *Saudi J. Biol. Sci.* 27, 2622–2631. doi: 10.1016/j.sjbs.2020.05.042

Jannink, J. L., and Gardner, S. W. (2005). Expanding the pool of PCR-based markers for oat. *Crop Sci.* 45, 2383–2387. doi: 10.2135/cropsci2005.0285

Jennings, T. N., Kaus, B. J., Mullis, T. D., Haig, S. M., and Cronn, R. C. (2011). Multiplexed microsatellite recovery using massively parallel sequencing. *Mol. Ecol. Resour.* 11, 1060–1067. doi: 10.1038/sj.hdy.6800939

Kalisz, S., and Kramer, E. (2008). Variation and constraint in plant evolution and development. *Heredity* 100, 171–177. doi: 10.1038/sj.hdy.6800939

Kassambara, A., and Mundt, F. (2020). Factoextra: extract and visualize the results of multivariate data analyses. R Package Version 1.0.7. Available at: https://CRAN.R-project.org/package=factoextra (Accessed December 12, 2023).

Kaur, G., Kapoor, R., Sharma, P., and Srivastava, P. (2021). Molecular characterization of oats (*Avena sativa* L.) germplasm with microsatellite markers. *Indian J. Genet. Plant Breed.* 81, 144–147. doi: 10.31742/IJGPB.81.1.18

Khan, A., Anjum, M. H., Rehman, M. K. U., Zaman, Q., and Ullah, R. (2014). Comparative study on quantitative and qualitative characters of different oat (*Avena sativa* L.) genotypes under agro-climatic conditions of Sargodha, Pakistan. *Am. J. Plant Sci.* 05, 3097–3103. doi: 10.4236/ajps.2014.520326

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., Haan, S., Engels, J. M. M., et al. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytol.* 233, 84–118. doi: 10.1111/nph.17733

Koide, Y., Sakaguchi, S., Uchiyama, T., Ota, Y., Tezuka, A., Nagano, A. J., et al. (2019). Genetic properties responsible for the transgressive segregation of days to heading in rice. *G3 Genes Genomes Genet.* 9, 1655–1662. doi: 10.1534/g3.119.201011

Koroluk, A., Paczos-Grzęda, E., Sowa, S., Boczkowska, M., and Toporowska, J. (2022). Diversity of Polish oat cultivars with a glance at breeding history and perspectives. *Agronomy* 12, 2423. doi: 10.3390/agronomy12102423

Koroluk, A., Sowa, S., Boczkowska, M., and Paczos-Grzęda, E. (2023). Utilizing genomics to characterize the common oat gene pool-the story of more than a century of Polish breeding. *Int. J. Mol. Sci.* 24 (7), 6547. doi: 10.3390/ijms24076547

Kozak, M., Bocianowski, J., Liersch, A., Tartanus, M., Bartkowiak-Broda, I., Piotto, F. A., et al. (2011). Genetic divergence is not the same as phenotypic divergence. *Mol. Breed.* 28, 277–280. doi: 10.1007/s11032-011-9583-9

Krishna, A., Ahmed, S., Pandey, H. C., and Bahukhandi, D. (2013). Estimates of Genetic variability, heritability and genetic advance of oat (*Avena sativa* L.) genotypes for grain and fodder yield. *Agric. Sci. Res. J.* 3, 56–61.

Kumar, R., Varghese, S., Jayaswal, D., Jayaswall, K., Yadav, K., Mishra, G., et al. (2023). Agro-morphological and genetic variability analysis in oat germplasms with special emphasis on food and feed. *PloS One* 18, e0280450. doi: 10.1371/journal.pone.0280450

Kumari, T., Jindal, Y., and Singh, S. (2017). Estimates of genetic variability, heritability and genetic advance in oats (*Avena* sp.) for seed and fodder yield traits. *Forage Res.* 43, 110–115.

Leišová-Svobodová, L., Michel, S., Tamm, I., Chourová, M., Janovska, D., and Grausgruber, H. (2019). Diversity and pre-breeding prospects for local adaptation in oat genetic resources. *Sustainability* 11, 6950. doi: 10.3390/su11246950

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: An r package for multivariate analysis. *J. Stat. Software* 25 (1). doi: 10.18637/jss.v025.i01

Li, C. D., Rossnagel, B. G., and Scoles, G. J. (2000). The development of oat microsatellite markers and their use in identifying relationships among *Avena* species and oat cultivars. *Theor. Appl. Genet.* 101, 1259–1268. doi: 10.1007/s001220051605

Liu, L., Guo, H., Zhao, L., Wang, J., Gu, Y., and Zhao, S. (2009). "Achievements and perspective of crop space breeding in China," in *Induced plant mutation in the genomics era*. Ed. Q. Y. Shu (Rome: Food and Agriculture Organization of the United Nations), 213–215.

Lyubimova, A. V., Tobolova, G. V., Eremin, D. I., and Loskutov, I. G. (2020). Dynamics of the genetic diversity of oat varieties in the Tyumen region at avenin-coding loci. *Vavilov J. Genet. Breed.* 24, 123–130. doi: 10.18699/VJ20.607

Mackay, I. J., Cockram, J., Howell, P., and Powell, W. (2021). Understanding the classics: the unifying concepts of transgressive segregation, inbreeding depression and heterosis and their central relevance for crop breeding. *Plant Biotechnol. J.* 19, 26–34. doi: 10.1111/pbi.13481

Mahdi, T., and Mohammad, R. (2022). Importance of BLUP method in plant breeding. *J. Plant Sci. Phytopathol.* 6, 040–042. doi: 10.29328/journal.jpsp.1001072

Majhi, P. (2020). "Heritability and its genetic worth for plant breeding," in *Advances in genetics and plant breeding*. Ed. P. Saidaiah (New Delhi, India: Akinik Publications), 69–75.

Maqbool, S. B., Zhong, H., Oraby, H. F., and Sticklen, M. B. (2009). "Transformation of oats and its application to improving osmotic stress tolerance," in *Transgenic wheat, barley and oats. Methods in molecular biology*. Eds. H. Jones and P. Shewry (New Jersey, USA: Humana Press). doi: 10.1007/978-1-59745-379-0_10

Mathias-Ramwell, M., Salvo-Garrido, H., Reyes-Rebolledo, M., and Montenegro-Barriga, A. (2016). Júpiter-INIA: a new oat variety with improved beta-glucan and protein contents. *Chil. J. Agric. Res.* 76, 401–408. doi: 10.4067/S0718-58392016000400002

Mazurkievicz, G., Ubert, I., de, P., Krause, F. A., and Nava, I. C. (2019). Phenotypic variation and heritability of heading date in hexaploid oat. *Crop Breed. Appl. Biotechnol.* 19, 436–443. doi: 10.1590/1984-70332019v19n4a61

Meira, D., Meier, C., Olivoto, T., Nardino, M., Klein, L. A., Moro, E. D., et al. (2019). Estimates of genetic parameters between and within black oat populations. *Bragantia* 78, 43–51. doi: 10.1590/1678-4499.2018116

Mohanta, T. K., Mishra, A. K., Mohanta, Y. K., and Al-Harrasi, A. (2021). Space breeding: the next-generation crops. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.771985

Montilla-Bascón, G., Sánchez-Martín, J., Rispail, N., Rubiales, D., Mur, L., Langdon, T., et al. (2013). Genetic diversity and population structure among oat cultivars and landraces. *Plant Mol. Biol. Rep.* 31, 1305–1314. doi: 10.1007/s11105-013-0598-8

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71.

Narváez, C., Castro, M. H., Valenzuela, B. J., and Hinrichsen, R. P. (2001). Patrones genéticos de los cultivares de vides más comúnmente usados en Chile basados en marcadores de microsatélites. *Agric. Téc.* 61 (3), 249–261. doi: 10.4067/S0365-28072001000300001

Nava, I. C., Duarte, I. T., de, L., Pacheco, M. T., and Federizzi, L. C. (2010). Genetic control of agronomic traits in an oat population of recombinant lines. *Crop Breed. Appl. Biotechnol.* 10, 305–311. doi: 10.1590/S1984-70332010000400004

Nersting, L. G., Andersen, S. B., von Bothmer, R., Gullord, M., and Jørgensen, R. B. (2006). Morphological and molecular diversity of Nordic oat through one hundred years of breeding. *Euphytica* 150, 327–337. doi: 10.1007/s10681-006-9116-5

Nouman-Khalid, M., Amjad, I., Vamuyah-Nyain, M., Sulyman-Saleem, M., Asif, M., Ammar, A., et al. (2021). A Review: TILLING technique strategy for cereal crop development. *IJACBS* 2 (5), 08–15.

ODEPA (2023) *Estadísticas productivas. Oficina de Estudios y Políticas Agropecuarias*. Available at: https://www.odepa.gob.cl/estadisticas-del-sector/estadisticas-productivas (Accessed July 18, 2023).

Olivoto, T., and Lúcio, A. D. (2020). Metan: An R package for multi-environment trial analysis. *Methods Ecol. Evol.* 11, 783–789. doi: 10.1111/2041-210X.13384

Olivoto, T., and Nardino, M. (2021). MGIDI: toward an effective multivariate selection in biological experiments. *Bioinformatics* 37, 1383–1389. doi: 10.1093/bioinformatics/btaa981

Paudel, D., Dhungana, B., Caffe, M., and Krishnan, P. (2021). A review of health-beneficial properties of oats. *Foods* 10, 2591. doi: 10.3390/foods10112591

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: Genetic analysis in excel. *Population Genet. Software Teach. Res. Mol. Ecol. Notes.* 6, 288–295. doi: 10.1111/j.1471-8286.2005.01155.x

Perry, M. C., and McIntosh, M. S. (1991). Geographical patterns of variation in the USDA soybean germplasm collection: I. morphological traits. *Crop Sci.* 31, 1350–1355. doi: 10.2135/cropsci1991.0011183X003100050054x

Premkumar, R., Nirmalakumari, A., and Anandakumar, C. R. (2017). Studies on genetic variability and character association among yield and yield attributing traits in oats (*Avena sativa* L.). *Int. J. Curr. Microbiol. Appl. Sci.* 6, 4075–4083. doi: 10.20546/ijcmas.2017.611.477

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Rana, M., Gupta, S., Kumar, N., Ranjan, R., Sah, R., Gajghate, R., et al. (2019). Genetic architecture and population structure of oat landraces (*Avena sativa* L.) using molecular and morphological descriptors. *Indian J. Tradit. Knowl.* 18, 439–450.

Raza, Q., Riaz, A., Saher, H., Bibi, A., Raza, M. A., Ali, S. S., et al. (2020). Grain Fe and Zn contents linked SSR markers based genetic diversity in rice. *PLoS One* 15, e0239739. doi: 10.1371/journal.pone.0239739

R Core Team (2022) *R: A language and environment for statistical computing*. Available at: https://www.R-project.org/.

Riaz, A., Kanwal, F., Ahmad, I., Ahmad, S., Farooq, A., Madsen, C. K., et al. (2022). New hope for genome editing in cultivated grasses: CRISPR variants and application. *Front. Genet.* 13. doi: 10.3389/fgene.2022.866121

Ristaino, J. B., Anderson, P. K., Bebber, D. P., Brauman, K. A., Cunniffe, N. J., Fedoroff, N. V., et al. (2021). The persistent threat of emerging plant disease pandemics to global food security. *Proc. Natl. Acad. Sci.* 118, e2022239118. doi: 10.1073/pnas.2022239118

Salgotra, R. K., and Chauhan, B. S. (2023). Genetic diversity, conservation, and utilization of plant genetic resources. *Genes* 14, 174. doi: 10.3390/genes14010174

Sanchez, D., Sadoun, S. B., Mary-Huard, T., Allier, A., Moreau, L., and Charcosset, A. (2023). Improving the use of plant genetic resources to sustain breeding programs' efficiency. *Proc. Natl. Acad. Sci.* 120, e2205780119. doi: 10.1073/pnas.2205780119

Šarhanová, P., Pfanzelt, S., Brandt, R., Himmelbach, A., and Blattner, F. (2018). SSR-seq: genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. *Ecol. Evol.* 8, 10817–10833. doi: 10.1002/ece3.4533

Silveira, S. F., da S. de C. S., Oliveira, D., Maltzhan, L. E., Corazza, T., de, V. F., Stulp, C., et al. (2020). Associations between agronomic performance and grain chemical traits in oat. *Commun. Plant Sci.* 10, 10.26814/cps2020001. doi: 10.26814/cps2020001

Skendžić, S., Zovko, M., Živković, I. P., Lešić, V., and Lemić, D. (2021). The impact of climate change on agricultural insect pests. *Insects* 12, 440. doi: 10.3390/insects12050440

Sürek, H., and Valentine, J. (1996). Relationship among some quantitative traits and heritabilities in cultivated oats (*Avena sativa* L.). *Tarim Bilim. Derg.* 2, 39–43.

Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., and Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci.* 61, 839–852. doi: 10.1002/csc2.20377

Szurman-Zubrzycka, M., Kurowska, M., Till, B. J., and Szarejko, I. (2023). Is it the end of TILLING era in plant science? *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1160695

Tanhuanpää, P., Manninen, O., Beattie, A., Eckstein, P., Scoles, G., Rossnagel, B., et al. (2012). An updated doubled haploid oat linkage map and QTL mapping of agronomic and grain quality traits from Canadian field trials. *Genome* 55, 289–301. doi: 10.1139/g2012-017

Tinker, N. A., Chao, S., Lazo, G. R., Oliver, R. E., Huang, Y., Poland, J. A., et al. (2014). A SNP genotyping array for hexaploid aot. *Plant Genome* 7, plantgenome2014.03.0010. doi: 10.3835/plantgenome2014.03.0010

Tinker, N. A., and Deyl, J. K. (2005). A curated internet database of oat pedigrees. *Crop Sci.* 45, 2269–2272. doi: 10.2135/cropsci2004.0687

Upadhyaya, H. D., Bramel, P. J., Ortiz, R., and Singh, S. (2002). Geographical patterns of diversity for morphological and agronomic traits in the groundnut germplasm collection. *Euphytica* 128, 191–204. doi: 10.1023/A:1020835419262

Vanjare, D. C., Shinde, G. C., Shinde, S. D., and Pawar, V. S. (2021). Genetic variability, heritability and genetic advance studies for green forage yield and associated traits in forage oat (*Avena sativa* L.). *Int. J. Curr. Microbiol. Appl. Sci.* 10, 488–493. doi: 10.20546/ijcmas.2021.1003.064

Viruel, J., Haguenauer, A., Juin, M., Mirleau, F., Bouteiller, D., Bouteiller, D., et al. (2018). Advances in genotyping microsatellite markers through sequencing and consequences of scoring methods for *Ceratonia siliqua* (Leguminosae). *Appl. Plant Sci.* 6 (12), e01201. doi: 10.1002/aps3.1201

Wang, L., Xu, J., Wang, H., Chen, T., You, E., Bian, H., et al. (2023). Population structure analysis and genome-wide association study of a hexaploid oat landrace and cultivar collection. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1131751

Weir, B. S. (1996). Genetic data analysis II: methods for discrete population genetic data. *Sinauer Sunderland Mass*, 445. p.

Wight, C. P., Yan, W., Fetch, J. M., Deyl, J., and Tinker, N. A. (2010). A set of new simple sequence repeat and avenin DNA markers suitable for mapping and fingerprinting studies in oat (*Avena* spp.). *Crop Sci.* 50, 1207–1218. doi: 10.2135/cropsci2009.09.0474

Winkler, L. R., Michael Bonman, J., Chao, S., Admassu Yimer, B., Bockelman, H., and Esvelt Klos, K. (2016). Population structure and genotype–phenotype associations in a collection of oat landraces and historic cultivars. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01077

Yan, W., Frégeau-Reid, J., Pageau, D., and Martin, R. (2016). Genotype-by-environment interaction and trait associations in two genetic populations of oat. *Crop Sci.* 56, 1136–1145. doi: 10.2135/cropsci2015.11.0678

Yemataw, Z., Blomme, G., Muzemil, S., and Tesfaye, K. (2018). Assessing qualitative and phenotypic trait diversity in Ethiopian enset [*Ensete ventricosum* (Welw.) Cheesman] landraces. *Fruits* 73, 310–327. doi: 10.17660/th2018/73.6.2

Zadoks, J. C., Chang, T. T., and Konzak, C. F. (1974). A decimal code for the growth stages of cereals. *Weed Res.* 14, 415–421. doi: 10.1111/j.1365-3180.1974.tb01084.x

Zheng, X., Cheng, T., Yang, L., Xu, J., Tang, J., Xie, K., et al. (2019). Genetic diversity and DNA fingerprints of three important aquatic vegetables by EST-SSR markers. *Sci. Rep.* 9, 14074. doi: 10.1038/s41598-019-50569-3

Zimmer, C. M., Ubert, I. P., Pacheco, M. T., and Federizzi, L. C. (2019). Variable expressivity and heritability of multiflorous spikelets in oat panicles. *Exp. Agric.* 55, 829–842. doi: 10.1017/S0014479718000418

Zohary, D., and Hopf, M. (2000). *Domestication of plants in the old world: the origin and spread of cultivated plants in West Asia, Europe, and the Nile Valley*. *3rd ed* (Oxford New York: Oxford University Press).

Zou, M., Tong, S., Zou, T., Wang, X., Wu, L., Wang, J., et al. (2023). New method for mutation inducing in rice by using DC electrophoresis bath and its mutagenic effects. *Sci. Rep.* 13, 6707. doi: 10.1038/s41598-023-33742-7

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 March 2024
DOI 10.3389/fpls.2023.1255506

OPEN ACCESS

EDITED BY
Patricio Hinrichsen,
Agricultural Research Institute, Chile

REVIEWED BY
Mohamed Cassim Mohamed Zakeel,
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Zaifeng Li,
Hebei Agricultural University, China

*CORRESPONDENCE
Casey Flay
caseyflay@gmail.com

RECEIVED 08 July 2023
ACCEPTED 25 October 2023
PUBLISHED 26 March 2024

CITATION
Flay C, Symonds VV, Storey R, Davy M and Datson P (2024) Mapping QTL associated with resistance to *Pseudomonas syringae* pv. *actinidiae* in kiwifruit (*Actinidia chinensis* var. *chinensis*).
Front. Plant Sci. 14:1255506.
doi: 10.3389/fpls.2023.1255506

Mapping QTL associated with resistance to *Pseudomonas syringae* pv. *actinidiae* in kiwifruit (*Actinidia chinensis* var. *chinensis*)

Casey Flay [1,2*], V. Vaughan Symonds[1], Roy Storey[2], Marcus Davy [2] and Paul Datson[2,3]

[1]School of Natural Sciences, Massey University, Palmerston North, New Zealand, [2]The New Zealand Institute for Plant and Food Research Limited, Te Puke, New Zealand, [3]Kiwifruit Breeding Centre, Te Puke, New Zealand

Pseudomonas syringae pv. *actinidiae* (Psa) is a bacterial pathogen of kiwifruit. This pathogen causes leaf-spotting, cane dieback, wilting, cankers (lesions), and in severe cases, plant death. Families of diploid *A. chinensis* seedlings grown in the field show a range of susceptibilities to the disease with up to 100% of seedlings in some families succumbing to Psa. But the effect of selection for field resistance to Psa on the alleles that remain in surviving seedlings has not been assessed. The objective of this work was to analyse, the effect of plant removal from Psa on the allele frequency of an incomplete-factorial-cross population. This population was founded using a range of genotypically distinct diploid *A. chinensis* var. *chinensis* parents to make 28 F_1 families. However, because of the diversity of these families, low numbers of surviving individuals, and a lack of samples from dead individuals, standard QTL mapping approaches were unlikely to yield good results. Instead, a modified bulk segregant analysis (BSA) overcame these drawbacks while reducing the costs of sampling and sample processing, and the complexity of data analysis. Because the method was modified, part one of this work was used to determine the signal strength required for a QTL to be detected with BSA. Once QTL detection accuracy was known, part two of this work analysed the 28 families from the incomplete-factorial-cross population that had multiple individuals removed due to Psa infection. Each family was assigned to one of eight bulks based on a single parent that contributed to the families. DNA was extracted in bulk by grinding sampled leaf discs together before DNA extraction. Each sample bulk was compared against a bulk made up of WGS data from the parents contributing to the sample bulk. The deviation in allele frequency from the expected allele frequency within surviving populations using the modified BSA method was able to identify 11 QTLs for Psa that were present in at least two analyses. The identification of these Psa resistance QTL will enable marker development to selectively breed for resistance to Psa in future kiwifruit breeding programs.

KEYWORDS

selection mapping, WGS, pool sequencing, QTLseqR, bulk segregant analysis

Introduction

Many cultivars of kiwifruit are devastated by the bacterial pathogen Psa (*Pseudomonas syringae* pv. *actinidiae* biovar 3), also known as the virulent form of Psa (Psa-V)(Everett et al., 2011; McCann et al., 2013; Dwiartama, 2017). This disease is particularly destructive to *A. chinensis* var. *chinensis* (Datson et al., 2015) genotypes, but also affects *A. chinensis* var. *deliciosa* (Takikawa et al., 1989). It has been reported that Psa spread from Asia, where up to four biovars were present (Koh et al., 1994). Each of the non-virulent biovars had different pathogenesis and molecular characteristics on different kiwifruit genotypes, but they did not cause the pathogenesis observed in the virulent Psa-biovar-3. This biovar, first reported in 2010 in New Zealand, is now widespread in the north island of the country, where kiwifruit is widely cultivated. Infected plants show symptoms such as leaf-spotting, cane dieback, wilting, or oozing a clear, brown or white liquid in spring or autumn from cankers (lesions). In highly susceptible genotypes, these symptoms occur on multiple canes leading to whole vine death. On more resistant genotypes, symptoms can involve flower bud browning, bud drop, flower wilting, and leaf spotting (Everett et al., 2011). To manage the outbreak, regulations were established by the national agency, Kiwifruit Vine Health (KVH), which required the removal of plants with severe symptoms such as cankers, or multiple dead canes. with severe symptoms such as cankers, or multiple dead canes, were removed.During the initial outbreak in 2011, the leading yellow-fleshed *A. chinensis* var. *chinensis* cultivar in New Zealand, named 'Hort16A', along with its pollenisers, were particularly susceptible to Psa. This susceptibility led to a significant decrease in gold kiwifruit production (Dwiartama, 2017). To address this decline, a gold-fleshed *A. chinensis* var. *chinensis* cultivar with the PVR name 'Zesy002' (fruit marketed as Zespri^{TM} SunGold) was utilised to replace gold fruit production that was previously reliant on 'Hort16A' (Everett et al., 2011; Dwiartama, 2017). This replacement cultivar exhibits greater resistance to Psa.

As Psa is such a ubiquitous and damaging pathogen, incorporation of resistance to Psa is required for any kiwifruit exposed to field conditions. Current breeding programmes that are based on crossing Psa-resistant families retain moderate resistance. However, due to the highly polygenic nature of resistance to Psa, strong resistance has not yet been achieved in the gold-fleshed *A. chinensis* var. *chinensis* (Tahir et al., 2018; Tahir et al., 2019). QTLs for resistance to Psa have been identified in two families of *A. chinensis* resulting from a resistant by susceptible cross, but identifying these QTL required large replicated trials from a single family and detailed phenotyping (Tahir et al., 2019). The phenotyping requirement, and the requirement of large, replicated families to generate QTL for resistance to Psa, could be overcome by identifying alleles remaining in breeding populations after the selective sweep caused by severe Psa infection.

When Psa spread within kiwifruit families established at Plant & Food Research, Kerikeri, New Zealand, notable differences in the extent of seedling removal emerged among these families due to Psa (personal communication Paul Datson).

The selective sweep caused by Psa presented a chance to investigate allele loci that persisted within the diverse families that made up this population. This investigation was carried out by using a technique called bulk segregant analysis (BSA), which aimed to gain insights into the genetic makeup related to resistance and susceptibility to Psa. BSA operates by assessing alterations in allele frequencies between populations that have segregated due to the pressures of selection, resembling a selection map (Michelmore et al., 1991; Wisser et al., 2008; Magwene et al., 2011; Li and Xu, 2022; Shen and Messer, 2022). The BSA technique has been used for the detection of QTL for target traits in various species, including dwarfing in watermelon (Dong et al., 2018), cotyledon colour in soybean (Song et al., 2017), cold resistance in rice (Sun et al., 2018), resistance to ascochyta blight in chickpea (Deokar et al., 2019), and kernel length-width ratio in wheat (Xin et al., 2020). A typical BSA investigates loci that differ between sample bulks segregating for a trait of interest, combining ideas from linkage mapping and GWAS (Michelmore et al., 1991; Li and Xu, 2022; Shen and Messer, 2022). Like classical linkage mapping, most BSA trials are designed using two parents with different phenotypes. The two parents are crossed to generate an F_1 population which is back-crossed or interbred for several generations to generate sufficient recombination to break up linkage from parents (Michelmore et al., 1991). Individuals from the last generation are selected to form two bulks that segregate for the phenotype of interest. Thus, alleles affecting the target phenotype should show a significant difference in frequency between the two bulks, while unselected alleles should remain in both bulks at similar frequencies (Michelmore et al., 1991; Shen and Messer, 2022). Diverging from the typical BSA, bulks can be analysed with BSA directly from F_1 populations (Dai et al., 2018; Guan et al., 2019). Similarly, selection mapping approaches compare a shift in allele frequency between two bulks created from samples of the population before and after a selection event altered the population's allele frequency (Wisser et al., 2008; Johnsson, 2018). The DNA that contributes to each of the bulks in BSA and selection mapping approaches is typically quantified for each individual, and an equivalent amount of DNA added to the bulk from each individual (Song et al., 2017; Dong et al., 2018; Munjal et al., 2018; Wang et al., 2021). While this approach ensures that a precise quantity of DNA is added from each individual, tracking samples and extracting DNA from individuals is costly and time-consuming. Moreover, both methods will include signal noise from a shift in allele frequency not caused by the target selection pressure. These unintended shifts in allele frequency can be caused by the genetics of the founding parents (James, 1970; Conolly et al., 2008; Chen et al., 2019).

An alternative approach to bulking DNA samples after extraction would be to bulk leaves of different individuals prior to DNA extraction. This approach would simplify sampling and reduce the cost and workload involved with DNA extraction by extracting DNA directly from a bulk of leaf samples. To help standardise the DNA contribution from each sample, the leaf sample growth stage and the amount of leaf material would need to be kept consistent. Samples from each individual then could be ground together for DNA extraction as a bulk. However, this approach precludes a precise balance of each individual's DNA contribution to the pool and may introduce greater variance into the bulks; therefore individuals that potentially contribute a greater

10.3389/fpls.2023.1255506

amount of DNA would make a greater contribution to the allele frequencies than others with less DNA extracted. This may decrease the power to identify allelic differences between bulks and thus QTL. Prior to applying such a modified method to an experimental population, a test of the approach to detect selection at a known site would need to be performed to determine its accuracy.

Testing the level of precision of the modified method of bulk sampling would require a population segregating for a simple control trait that is determined by a single well-characterised locus, and ideally with low interaction between the gene and the environment. To this end, within the dioecious *A. chinensis*, plant sex is a suitable trait as it is easy to phenotype and it is controlled by a single well-characterised dominant gene that is not affected by the environment (Akagi et al., 2018). This kiwifruit sex gene, named *Shy Girl* (*SyGI*) exists on the male Y chromosome and suppresses flower feminisation, producing males in plants possessing it (Akagi et al., 2018). It was assumed that if a shift in allele frequency could be detected using these methods in the monogenic *Shy Girl* gene, polygenic loci of strong influence on the population would also be able to be detected.

This study aimed first to test whether pooling leaves from multiple individuals prior to DNA extraction enables a BSA to be effectively carried out on the resulting DNA pool. This was done using a series of bulked pools that varied in the ratio of male and female *A. chinensis* individuals that contributed to the pools and investigated whether the QTL for plant sex could be identified on chromosome 25. Part two of this work aimed to identify any changes in allele frequency between bulks of sample pools of seedlings that had survived in the field following a Psa selective sweep and a bioinformatically generated bulk of data from parents contributing to each sample bulk. Regions of the genomes where alleles have a greater sample depth in the WGS of sample bulk data than expected from their parental bulk of data should highlight the regions of the genome under strong selection from Psa.

Materials and methods

Population

A diverse population of diploid *A. chinensis* var. *chinensis*, named "12x18", was identified at Plant & Food Research, Kerikeri, New Zealand as a suitable population to meet the objectives of both aspects of this study. The parental seedling vines for this population were initially planted in 2015 and cultivated using a T-bar system, with a spacing of 0.75 m between each plant and 3 m between rows. The population was strategically distributed across three blocks, each spanning 4000 m^2, with 6-m high hedging shelter belts serving as dividers and boundaries around the blocks. This population was naturally exposed to Psa, which was present in the Kerikeri orchard at the time of seedling planting. The exposure to Psa led to the development of symptoms in certain individuals including tip dieback, cane death, oozing from infected cankers, and in highly susceptible cases, complete plant death. To manage Psa symptoms, canes were removed if tip death or cane death was observed on a single cane. When more than three canes were infected with cankers or experienced cane dieback the entire vine was removed from the orchard. The structure of the 12x18 population was established by the crossing a diverse set of 12 female and 18 male parents from *Actinidia* germplasm, employing an incomplete factorial design that resulted in 63 families. A variable number of seedlings (33, 48, or 56) from each family were planted in the field after their initial establishment in pots. Fifty-nine families had individuals remaining after 4 years (Figure 1, Table 1, Figure 2). Between 2015 and 2019, severe Psa infections led to the removal of individuals from various families with a range from 63% to 100%. Of the families that had surviving individuals, only 25 had sufficient a sufficient number of individuals to be included in the current study. However, it was necessary to pool families based on their parentage due to the relatively low numbers in individual families.

Female \ Male	P1	P2	P3	P4	P5	P6	P7	P8
P9		6% (3/48)		15% (7/48)		54% (26/48)	17% (8/48)	
P10	4% (2/56)				13% (7/56)			27% (9/33)
P11			27% (13/48)	44% (21/48)				
P12	63% (30/48)	19% (9/48)				50% (24/48)		
P13			8% (4/48)					
P14				4% (2/56)				
P15				2% (1/48)				
P16		11% (6/56)						
P17		7% (4/56)						
P18	2% (1/56)		38% (18/48)					
P19		6% (3/48)		13% (6/48)				
P20			4% (2/48)					
P21	6% (3/48)							
P22	4% (2/48)	23% (11/48)						

FIGURE 1
The crossing structure of 25 families with surviving individuals from the 12x18 population. For part two of this work, eight bulks were made by sampling leaves from all surviving plants within families which shared the parent indicated in the blue columns and salmon rows. Female parents (P9-P22) are shown on the left-hand side with male parents (P1-P8) at the top. Percentages in cells indicate the number of individual F_1 seedlings remaining after being exposed to Psa (*Pseudomonas syringae* pv. *actinidiae*) for four years in the field. The numbers in brackets are the number of individuals remaining from the total number of individuals that were planted from the family. For example, 4%, or two of 56 plants survived after four years in the field from the cross between P1 and P10.

10.3389/fpls.2023.1255506

TABLE 1 The number of males and females contributing to bulks in part one of this work.

Bulk	Number of males in bulk	Number of females in bulk	Total individuals in bulk	Percentage of males in bulk	Percentage of females in bulk
1	2	17	19	10.5	89.5
2	4	15	19	21.1	78.9
3	6	13	19	31.6	68.4
4	9	11	20	45	55
5	10	10	20	50	50
6	12	8	20	60	40
7	15	5	20	75	25
8	16	4	20	80	20
9	18	2	20	90	10

The percentage of males within each of the nine bulks ranged from 10.5% to 90%.

Sample collection, DNA extraction and sequencing

The field sampling, DNA extraction processes, and sequencing methods were the same for both parts of this work. Sampling of plant material was done by placing a single leaf from each plant, destined for a bulk, into a plastic bag labelled with the bulk's name. Leaf samples were taken from the third leaf from the growing cane tip and kept cold in a chilly box with ice while sampling. After field collection, the bulks of leaves were stored in a -80°C freezer before processing. A 10-mm diameter leaf disc was collected from the lamina of each leaf while frozen. All leaf discs from a bulk were finely ground together in liquid nitrogen with a pestle and mortar. DNA was extracted from the ground material with a Qiagen DNeasy® Plant Maxi kit. To remove pectin from samples, DNA was precipitated by adding 1/10 volume sodium acetate (3 M, pH

			Sample bulk parent (male)				Sample bulk parent (female)			
	Parents with WGS	bulk parent crossed with	P1	P2	P3	P4	P9	P10	P11	P12
Male polleniser of female sample bulk parent	WGS	P1						5.6%		23.8%
	WGS	P2					3.4%			7.1%
	WGS	P3							19.1%	
	WGS	P4					8.0%		30.9%	
	noWGS	P5						19.4%		
	noWGS	P6					29.5%			19.0%
	noWGS	P7					9.1%			
	WGS	P8						25.0%		
Female parent used in cross with male sample bulk parent	WGS	P9		4.2%		9.5%				
	WGS	P10	2.6%							
	WGS	P11			18.1%	28.4%				
	WGS	P12	39.5%	12.5%						
	WGS	P13			5.4%					
	noWGS	P14				2.7%				
	WGS	P15				1.4%				
	noWGS	P16		8.3%						
	noWGS	P17		5.6%						
	noWGS	P18	1.3%		24.3%					
	noWGS	P19		4.2%		8.1%				
	noWGS	P20			2.7%					
	WGS	P21	3.9%							
	noWGS	P22	2.6%	15.3%						
Expected percentage of sample bulk DNA contributed by parents samples were bulked on			50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%
Expected percentage of sample bulk DNA contributed by parents with WGS data			96%	67%	73%	89%	61%	81%	100%	81%
Expected percentage of sample bulk DNA missing from parent bulk			4%	33%	27%	11%	39%	19%	0%	19%
Bulk name			B1	B2	B3	B4	B9	B10	B11	B12

FIGURE 2
Percentage of each parent's theoretical contribution to sample bulks. Bulks were based on the parents in columns, with parents contributing to the sample bulk in rows. Families from parents with grey-filled parent names were represented twice where the family was used in bulks based on male and female sample bulk parents. Parents with whole-genome sequence (WGS) data have cells filled in green, and those without WGS are filled in salmon. Parent bulks contained DNA only from the parents with green shading. The total theoretical DNA contribution missing from parents without WGS data in parent bulks is shown in the bottom row.

5.2) to two times the volume (calculated after addition of sodium acetate) of at least 95% ethanol. Samples were incubated on ice overnight, then centrifuged at 14000 g for 30 min at 4° C. Supernatant was removed and rinsed with 70% ethanol, then centrifuged at 14000 g for 15 min. The supernatant was discarded, and the pellet dissolved in TE buffer (pH 8.0). TE buffer was made by adding 100 mL of 1M Tris-Cl (pH 8.0) to 20 uL of 0.5 M EDTA (pH 8.0) to 9.880 mL of reverse-osmosis water. DNA quality and quantity were checked using a Qubit® 2.0. In samples with a low DNA quantity, extraction was repeated. In samples with low-quality DNA, identified by a 260/280 value of under 1.6, DNA was cleaned of pectin using a second ethanol precipitation step. In this step, DNA was precipitated in 98% ethanol and the DNA pellet was lightly massaged with a spatula against an Eppendorf tube wall to remove pectin within the DNA precipitate. A minimum of 1400 ng of DNA from each bulk was sent to the Australian Genome Research Facility (AGRF) for PCR free library preparation and whole-genome sequencing at 30x coverage with 150 bp paired-end reads using the Illumina NovaSeq 6000 platform.

Using this method of bulking leaf samples forgoes the step of extracting DNA from individual plants, quantifying the DNA from each extraction and adjusting the amounts so that each bulk contains an even amount from each contributing individual. However, it also introduces the risk of having a variable quantity of DNA added from each individual to a bulk and may therefore increase the error associated with analysing allele depth.

Comparing the genomic difference between the parents contributing to bulks

Because the methods used in part two of this work could be influenced by the similarity of parents, the genomic distance between parents needed to be tested. The genomic distance between individuals can be analysed with principal component analysis by transforming genomic data into a Boolean vector, as described in Konishi et al. (2019). The variants from parents were used to identify the genomic distance between each parent with whole genome sequence (WGS) data.

Bulking samples for part one of this study

The sensitivity of the sampling and bulk segregant analysis methods to detect a shift in allele frequency was investigated. This was achieved by using male and female F_1 individuals from a mix of families from the 12x18 population that had parent P8 as the father (Figure 1, Figure 2). Female parents of these families included P10, P14, P15, and P17. The shift in allele frequency at the sex locus on chromosome 25 was tested between the nine bulks of DNA containing about 10%, 20%, 25%, 40%, 50%, 55%, 68.4%, 78.9% or 89.5% male contribution to the bulk from 19–20 individuals (Table 1).

Bulking samples for part two of this study

Part two of this work investigated WGS data from bulks to detect whether there was a shift in allele frequencies within bulks of individuals that remained in families after exposure to Psa. This work was complicated because severe Psa infection had led to the removal of many individuals from all the families in the population. Because some families had very few individuals remaining, each of the eight sample bulks included resistant individuals from up to six families. These were bulked based on a single parent that contributed to all the families in the bulk (Figure 1, Figure 2). For example, the bulk B1 contained F_1 families from crosses P1 x P10, P1 x P12, P1 x P18, P1 x P21, and P1 x P22.

Part two of this work differed from a typical BSA because there was no DNA from individuals that were removed because of Psa. Instead, the frequency of alleles in surviving individuals was compared against a bioinformatically generated bulk of data from parents contributing to the sample bulk. The bioinformatically generated parent bulks were used in place of bulks of individuals susceptible to Psa. Bulks like this can be used because the alleles in the parental bulks were representative of the families included in the bulks without selection. This methodology is similar to that done for selection mapping (Wisser et al., 2008; Matsumoto et al., 2017), but it has the drawback of assuming no other influences on allele transmission. The bioinformatically generated parent bulks were made by merging parental BAM files before variant call files (VCF) were made. However, not all parents that contributed to the families used in parental bulks had WGS data available (Figure 2, Table 2). As a result, the bulks of parents that contained some parents without WGS data would give a less accurate representation of the population before selection. The loss of information was particularly apparent in the bulk of B9, which had 29.5% of its theoretical DNA contribution missing from its P6 parent and 9.1% missing from the P7 parent (Figure 2). The missing data from parents would have resulted in some alternate alleles present in these parents not being included in the analysis. Unfortunately, once the pools were established during field sampling, the families with missing WGS data could not be removed from pools.

Bulk segregant analysis part one

To test the limits of the methods used to bulk samples and extract DNA to determine the architecture for Psa tolerance, WGS data for part one of this work were analysed with the QTLseqR package v0.7.5.2 (Mansfeld and Grumet, 2018). The analysis included nine bulks, with a varying number of males added to each bulk, were each compared with each other for 36 separate analyses, described further below. These comparisons were expected to present a QTL peak in the bulk segregant analysis at 1.6 Mb on chromosome 25. QTLs were expected on chromosome 25 because it contains the heterozygous dominant sex-determining *Shy Girl* gene that suppresses the feminisation of flower production to generate male flowers and thus a male plant (Akagi et al., 2018).

TABLE 2 Depth filter settings applied to data before Gprime analysis.

Sample bulk	Refence allele frequency	Minimum total depth	Maximum total depth	Depth difference	Minimum sample depth	Minimum genomic quality
B1	0.05	40	85	50	10	100
B2	0.05	35	85	50	10	100
B3	0.05	20	85	50	10	100
B4	0.05	40	150	50	10	100
B9	0.05	40	130	50	10	100
B10	0.05	40	90	50	10	100
B11	0.05	60	170	50	10	100
B12	0.05	30	100	50	10	100

Sample bulks B1-B12 retained the same reference allele frequency, allele depth difference, minimum sample depth and minimum genomic quality, but varied in the minimum total allele depth and maximum total allele depth depending on the distribution of depth data in each sample bulk.

But detection of QTL at the *Shy Girl* gene locus could only occur if the methods used were tolerant enough of the sampling and bulking methods, the effect of Psa on the families, and the relationship between the samples for the bulks, since these would have an influence on frequency of alleles between bulks. For example, if bulks with a 5% difference in male number were compared and QTL were consistently detected in comparisons with different backgrounds, and with a similar difference in male percentage between pools, it could be assumed that the methodology added 5% of error to the analysis.

Bulks of males were compared with each other because WGS data were unavailable for two of the five male bulk parents, P14 and P17. Thus, a bulk of these parents would not accurately represent the bulks of parental data. Instead, data from each of the nine bulks with a known percentage of males were compared with each other, resulting in 36 separate analyses that compared pairs of bulks. The difference between the percentages of males between bulk pairs varied between 5 and 79.5%.

Binary alignment files (BAM files) of WGS data were generated from compressed FASTQ formatted sequence files containing single read sequence output by aligning reads to an unpublished in-house reference genome of parent P8 by Roy Storey using BWA-MEM (Yao et al., 2020). The author completed the subsequent bioinformatics analyses using the R coding language. BAM files from separate flow cell lanes were merged with Picard "MergeSamFiles". Samtools was used for sorting and indexing BAM files. Variant call files were generated using BCFtools mpileup with options including setting a minimum base quality of 20 and disabling probabilistic realignment to help reduce false SNPs caused by misalignments. Indel calls were excluded. Optional tags included the depth at each site, the depth of each allele, and the Phred-scaled strand bias P-value. Uncompressed output was piped to BCFtools call, which included the genomic quality and genotype posterior probability format fields and the multiallelic caller option. The resulting variant call files were indexed using BCFtools index. BCFtools query was used to split data into separate comma-separated value text files for each chromosome and exclude sites with a depth of less than 20 or greater than 200, and data were read into R/datatable.

The SNP index for each bulk pair to be analysed was calculated by dividing the alternate allele depth by the total read depth. The reference allele frequency was calculated by summing the reference allele depth of bulks being compared and dividing the result by the sum of the total depth of the bulks being compared. The delta-SNP index was calculated by subtracting the SNP index of the sample bulk from the parental bulk. The modified G statistic was calculated for each SNP based on the observed and expected allele depths (Magwene et al., 2011) and smoothed using a tricube smoothing kernel (Watson, 1964) in QTLseqR (Mansfeld and Grumet, 2018). The Gprime value was calculated from the tricube smoothed G statistic by taking the average weight of the physical distance across the neighbouring SNPs within the 1-Mb window. This approach accounted for linkage disequilibrium and minimized the noise attributed to SNP calling errors (Magwene et al., 2011). SNPs were filtered using the QTLseqR package selecting a reference allele frequency of 0.05, a minimum total depth of 60, a maximum total depth of 160, an allele depth difference of less than 50 between bulks, a minimum sample depth of 10 and a minimum genomic quality of 100. The QTLseqR analysis package had the bulk size set to 20 individuals with the Gprime window size set at 1 Mb. Because the adjusted p statistic threshold failed to detect peaks with a low difference in allele frequency, the genomic position of the top 0.5% of SNPs were used as the peak locus.

Bulk segregant analysis part two

Part two of this work used the same bioinformatics pipeline as part one of this work to generate VCF files of sample bulks. However, part two of this work differed from part one because a modified BSA approach was used for bulk creation and sample bulks were compared against the bioinformatically generated bulk of WGS data from parents that contributed to the bulked families (Table 1). Bulks were selected for analyses based on the presence of WGS data for parents and having greater than ten individuals available for sampling. These parent bulks were created by merging BAM files from parents using samtools-merge. VCF files were

created with BCFtools-mpileup using the same options as in part one of this work.

Before the sample bulks and parent bulks could be compared, calculations based on SNP data from VCF files were performed. VCF files were read into R/datatable, and a SNP index was calculated for each bulk by dividing the alternate allele depth by the total read depth. The reference allele frequency was calculated by summing the reference allele depth of the sample bulk and the parent bulk, and dividing the result by the sum of the total depth of the bulks being compared (Mansfeld and Grumet, 2018). The delta-SNP index was calculated by subtracting the SNP index of the sample bulk from the parental bulk (Mansfeld and Grumet, 2018).

The data preparation for analysis in QTLseqR was done similarly to that done in part one. First, BCFtools-query split data into separate.csv files for each chromosome and excluded sites with a depth of less than 20 or greater than 200. The SNP index per bulk was calculated by dividing the alternate allele depth by the total read depth. Unlike in part one, in part two the reference allele frequency was calculated using the sum of reference allele depths of sample bulks and the result was divided by the sum of the total depth of the parental bulks. The delta-SNP index was calculated by subtracting the SNP index of the sample bulk from the parental bulk.

Data from each of the eight sample bulks were compared with their parent bulk using the Gprime analysis portion of the QTLseqR package (Mansfeld and Grumet, 2018). Gprime analysis was used because the average G values across SNPs in the 1-Mb sliding window reveal the signal of divergence in allele frequency between bulks that are conserved between closely linked sites (Magwene et al., 2011; Mansfeld and Grumet, 2018). Using the G value reduces the influence of random noise due to variable sequencing read coverage (Mansfeld and Grumet, 2018). Within the QTLseqR package, SNPs were filtered by depth for each comparison depending on the data distribution. Minimum and maximum total depth were set to remove SNPs of extremely low and extremely high frequency (Table 2) (Mansfeld and Grumet, 2018). Filtering SNPs by read depth helps remove SNPs with low confidence due to low coverage, or remove SNPs in repetitive regions that would have an artificially inflated read depth (Mansfeld and Grumet, 2018). Settings for the Gprime analysis method were as follows: the sliding window size was set at 1 Mb, the outlier filter was set as "deltaSNP", and the filter threshold was set at 0.4. The resulting Gprime values for each SNP site were plotted with the top 0.5% of Gprime values and peak loci for each bulk plotted in Figure 3.

Results

The PCA comparing the parents that contributed to bulks that also had WGS data showed a close relationship among the half-sibling individuals P8 and P9, with a greater distance between P8 and P9 and the other individuals at PC1. PC2 showed an even distribution of genomic relationship between the remaining individuals with the exception of P1 and P2, which had minimal genomic distance on PC2 (Figure 4).

The methods used for the BSA in this work differed from the standard methods used for BSA in QTLseqR (Mansfeld and Grumet, 2018). Therefore, the sensitivity of these methods to detect QTL in bulks of *A. chinensis* var. *chinensis* needed to be tested. Part one of this work tested the resolving power of the methods by making 36 pairwise comparisons at the sex loci on chromosome 25 among the nine bulks of individuals with a differing percentage of males. However, QTLs were detected in only 12 of the 36 bulk comparisons when using the adjusted p =

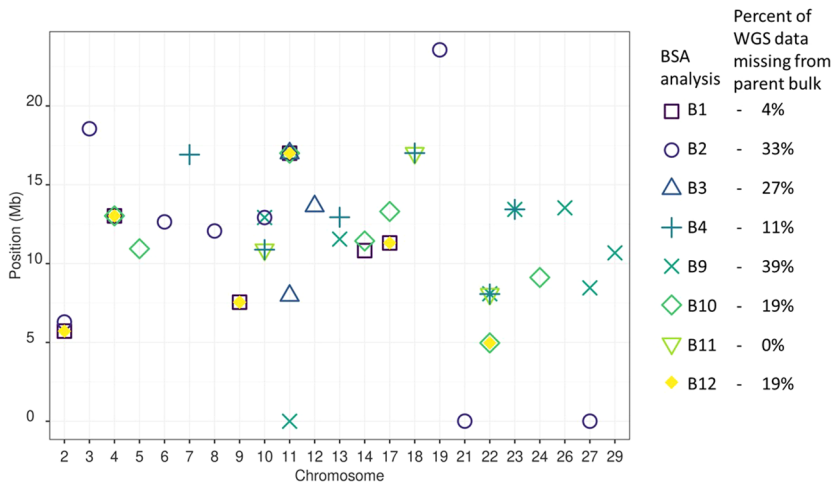


FIGURE 3
BSA QTL peak positions for Psa resistance between all analyses. Thirty QTLs were detected among the eight bulks analysed. The BSA presenting the most unique QTL were B2, B9, and B10. The BSA presenting no unique QTL were B11 and B12. A single QTL site was common between four BSA on Chromosome 11 at 16.95 Mb from B1, B3, B10 and B12. Four QTL sites were common between three BSA on chromosome 2 at 5.35 Mb, chromosome 4 at 13.02 Mb, Chromosome 17 at 11.31 Mb, and Chromosome 22 at 8.07 Mb. Six QTL sites were found in common between two BSA on Chromosome 9 at 7.55 Mb, Chromosome 10 at 10.89 and 12.92 Mb, Chromosome 14 at 11.13 Mb, Chromosome 17 at 11.31 Mb, Chromosome 18 at 17.01 Mb, Chromosome 22 at 4.97 Mb, and Chromosome 23 at 13.44 Mb. The 17 remaining QTL sites were found in a single BSA.

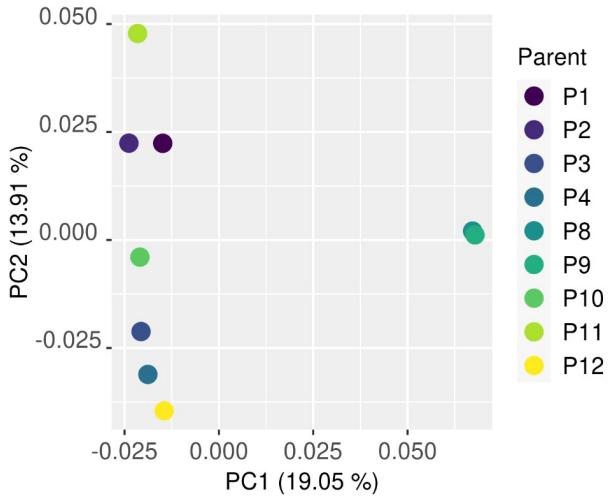


FIGURE 4
The genomic distance between parents that contributed to bulks analysed by principal component analysis. A close relationship among the half-siblings P8 and P9 was found with a greater distance between P8 and P9 and the other individuals at PC1. The close relationship between P8 and P9 remained at PC2, with a close relationship between P1 and P2.

0.05 threshold (Figure 5). Increasing the threshold to adjusted p = 0.1 included more QTL peaks, but also significantly increased the signal-to-noise ratio. Because the adjusted p-value based threshold could be caused by the alternative method of bulking multiple families or lack of inclusion of some parents in the bioinformatically generated bulk of parents, the significance threshold was changed to use the top 0.5% of Gprime values. Using the top 0.5% of Gprime values allowed QTL detection from BSA with greater accuracy in bulks, detecting QTL at the sex-linked gene locus in 19 of the 36 bulks analysed. However, using the top 0.5% of Gprime values as a threshold of significance for QTL detection also has the disadvantage of missing smaller peaks for QTL in BSA plots where the signal for selection for some QTL was strong and covered a large range of loci. For example, the P3 bulk may have signal for selection on chromosomes 25 and 29, but the peaks on chromosomes 11 and 12 hold the top 0.5% of Gprime values.

To determine the effect of Psa on *A. chinensis* var. *chinensis* alleles in an incomplete factorial population, in part two of this work, samples that survived Psa were assigned to bulks based on families with a parent in common (Figure 1, Figure 2). Using BSA, the eight sample bulks were compared against bioinformatically created bulks of parental WGS data. The eight resulting BSA, presented in Supplementary Material, identified sites of higher frequency in the sample bulk compared to the parent bulk, potentially caused by selection for resistance to Psa. The QTL presented as higher Gprime values in the resulting BSA plots, with the top 0.5% of Gprime values considered significant QTL. These significant QTL were summarised between bulks in Figure 3.

In theory, the variants for resistance to Psa had a maximum potential selection of 50%. For example, in the case of a cross ab x cc with resistance associated with the 'a' variant, if there was strong selection for the 'a' variant in all seedlings, the resulting family containing ac variants would have a variant frequency 50% higher than if there was no selection producing ac and bc variants. The b variant will also decrease in frequency by 50%.

Discussion

Psa is one of the most destructive diseases affecting kiwifruit, with a broad range of susceptible and tolerant *A. chinensis* var. *chinensis* genotypes. However, QTLs for Psa resistance have been published within only two families to date (Tahir et al., 2019; Tahir et al., 2020), potentially leaving many alleles for resistance to Psa undiscovered. Typically, QTL mapping methods would be used to investigate loci for a polygenic trait such as Psa (Jansen, 1996; Lefebvre and Palloix, 1996). However, accurately identifying traits influenced by more than one locus with QTL mapping is a costly and resource-intensive process requiring large replicated families specially developed for this purpose (Wisser et al., 2008; Soto-Cerda and Cloutier, 2012; Tahir et al., 2018; Gupta et al., 2019; Tahir et al., 2019). The work presented here overcame the limitations of the typical QTL mapping process by using bulks of diverse F_1 families in a modified BSA. This approach further increased the utility and cost-effectiveness of the typical BSA methods by analysing multiple small families in a single bulk, decreasing the sampling complexity, reducing the DNA extraction time and cost, reducing sequencing costs, and increasing the breadth of Psa resistance alleles that could be detected within a bulk.

Bulked DNA has been analysed using BSA methods for a wide range of species and traits (Michelmore et al., 1991; Song et al., 2017; Dong et al., 2018; Sun et al., 2018; Xin et al., 2020; Li and Xu, 2022; Shen and Messer, 2022). The approach taken here used BSA analysis methods to identify alleles for Psa resistance by measuring the shift in allele frequency of an *A. chinensis* var. *chinensis* population that had many individuals removed from established

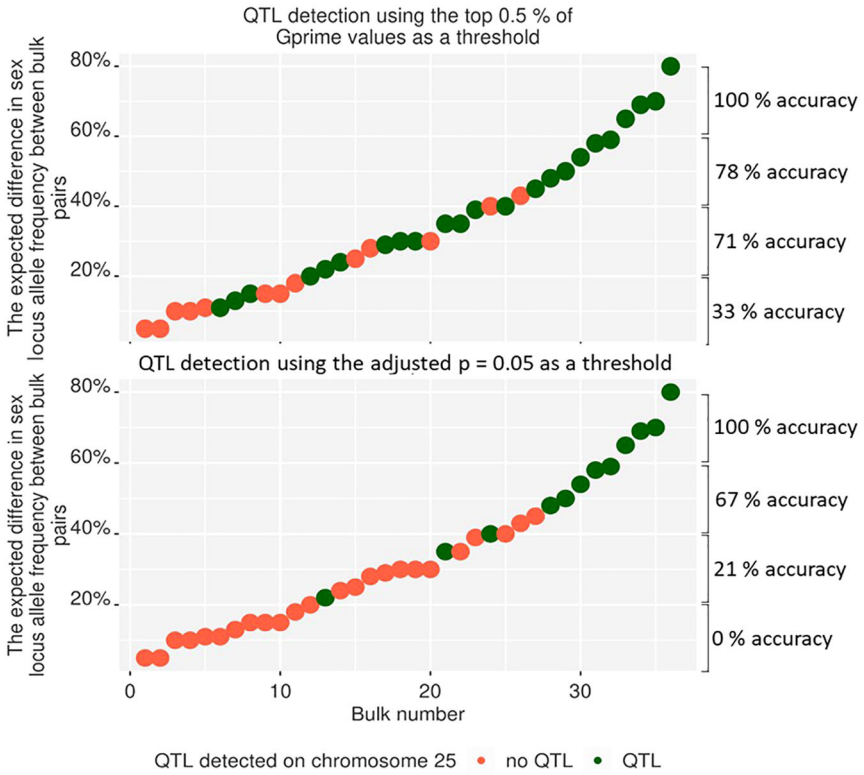


FIGURE 5
The detection of QTLs on chromosome 25 from pairwise comparisons between nine bulks with a different percentage of males added to each bulk. This figure presents only QTL from chromosome 25, but the analysis was completed over the whole genome. Using the top 0.5% of SNPs over the whole genome (top) to detect QTL peaks gave fewer false-negative results than QTL peaks, which were deemed statistically significant using an adjusted p-value of greater than 0.05. When using the top 0.5% threshold, the detection accuracy was estimated to be 33%, 71%, 78%, and 100% for an expected difference in allele frequency between bulks of 0–20%, 20–40%, 40–60%, and 60–80%, respectively. When using the adjusted p=0.05 threshold, the detection accuracy was estimated to be 0%, 21%, 67%, and 100% for an expected difference in allele frequency between bulks of 0–20%, 20–40%, 40–60%, and 60–80%, respectively.

families due to Psa. The selective sweep of susceptible individuals provided a prime opportunity to measure the effect of this selection on the alleles that remained within those *A. chinensis* var. *chinensis* families. However, because susceptible plants were not sampled prior to loss, DNA was not captured from susceptible individuals, which a typical BSA would use as the comparison bulk (Michelmore et al., 1991; Li and Xu, 2022). The lack of a susceptible bulk was compensated for in this modified BSA by using a bulk of parent WGS data as a population not exposed to selection pressure for Psa resistance. To our knowledge this is the first time a bulk of parents has been used as a substitute for a bulk in a BSA analysis and the first time QTL for kiwifruit resistance to Psa have been identified within a diverse range of families.

Testing the sensitivity of the modified methods

Part one of this work tested the ability of the modified BSA method to analyse the effects of Psa in kiwifruit by identifying a known locus through detecting a shift in allele frequency between bulks. This test was required because the families were sampled and extracted as bulks in a unique way (Figure 1). The modified method involved sampling leaves for a bulk from the field into a single bag and extracting DNA from all leaf samples within the bulk by first grinding them all together before DNA extraction. This significantly increased the speed of sampling, sample tracking and DNA extraction, but variance among leaves could still occur due to differences in the number and size of cells, ratio of mitotic to interphase nuclei, or differing structures or biochemical composition of the plant cells (Marsal et al., 2013). As a result, changes in allele frequency could have been created between samples due to a variable amount of DNA extracted from each leaf sample. Testing the effect of the sampling methods and their integration with the analysis methods on allele detection found that the detection of an allele frequency shift of over 10% was effective for detecting alleles under strong selection for Psa resistance for part two of this work. The detection accuracy was enhanced by modifying the threshold of detection for QTL. The modified thresholds allowed the detection of a 20–40% shift in male allele frequency with an accuracy above 71% (Figure 5), and a shift in allele frequency above 60% could be detected with 100% accuracy. This confirmed that the modified technique could detect large changes in allele frequency. Because strongly selected alleles can cause a shift in allele frequency of up to 50% within a family (Miko, 2008), this increased the confidence that sites of strong selection

could be detected in part two of this work. However, these methods may not detect small shifts in allele frequency.

QTL for resistance to Psa

A typical BSA groups together loci that differ in allele frequency between sample bulks segregating for a trait of interest (Michelmore et al., 1991; Li and Xu, 2022). However, part two of the approach taken here modified the bulking strategy of a typical BSA. Each sample bulk was made up of multiple diverse families that had DNA extracted from all individuals in the bulk simultaneously as a single sample. Part two of this work also differed from a typical BSA because no individuals culled due to Psa infections were sampled to make a comparison bulk. Instead, a bulk of parental alleles was bioinformatically generated using WGS data from parents of the families under investigation. The alleles with a higher frequency in the sample bulk compared to the parent bulk were assessed against the significance thresholds established in the preliminary trial to identify alleles for Psa resistance. Between the eight sample bulks that each contained three to seven families, 30 QTLs for resistance to Psa were identified. Twelve of the 30 QTLs were detected in more than one bulk, with one locus on Chromosome 11 at 16.95 Mb detected among four bulks (Figure 6).

Alleles for Psa resistance have previously been published in a large diploid *A. chinensis* var. *chinensis* family (Tahir et al., 2019). That study identified two QTL using field scores for Psa resistance. One of the loci identified was in an identical location to the locus found in this work on Chromosome 22 at 4.967790 Mb from two bulks, B10 and B12. Because the parent P8 was used in the bi-parental population by Tahir et al. (2019), and bulk B10 from this work also contained the parent P8, it seems likely that this locus for resistance to Psa is coming from parent P8. The other locus for Psa resistance detected by Tahir et al. (2019) using field scores, found on Chromosome 27 at 4.305319 Mb, was not detected in this study. The lack of detection of this locus may have been because the cultivar 'Hort16A' that identified as the parent that contributed Psa resistance to the bi-parental family was not included in this study. A parent of 'Hort16A', included in this study, named parent P13, was included in bulk B3 but the contribution of P13 to this bulk was low at 5.4%. It would be expected to have a peak if the resistance allele was strongly selected for in the B3 bulk, but it is more likely that the other parent of 'Hort16A', named CK15_01 by Tahir et al. (2019), was the contributor of the resistance QTL found by Tahir et al. (2019) on Chromosome 27 at 4.305319 Mb.

The commonality of resistance allele sites among some of the different BSA bulks may reflect the inclusion of common parents that contributed to those bulks. This commonality of QTL sites can give insight into which of the parents were likely to be contributing the alleles under selection at some of the QTL. Looking at the peaks that are at the same site on the same chromosomes in different bulks (Figure 6) allows us to infer the most likely parents that were contributing the alleles to those QTL. For example, the QTL on Chromosome 2, at 5.35 Mb, is shared between bulks B1, B2, and B12, indicating that parent P12 is likely to be the source of alleles in higher frequency in those bulks (Figure 6). Similarly, the parent P4 is likely the source of the QTL on Chromosome 22 at 8.07 Mb and the parent P1 is likely the source of the QTL on Chromosome 4 at

BSA QTL chromosome and position	Bulk name	Parents contributing to bulks																					
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
Chr. 2, 5.35 Mb	B1	P1									P10		P12						P18			P21	P22
	B2		P2							P9			P12				P16	P17		P19			P22
	B12	P1	P2				P6						P12										
Chr. 4, 13.02 Mb	B1	P1									P10		P12						P18			P21	P22
	B10	P1				P5			P8		P10												
	B12	P1	P2				P6						P12										
Chr. 9, 7.55 Mb	B1	P1									P10		P12						P18			P21	P22
	B12	P1	P2				P6						P12										
Chr. 10, 10.89 Mb	B4				P4					P9		P11			P14	P15				P19			
	B11			P3	P4							P11											
Chr. 10, 12.92 Mb	B2		P2							P9			P12				P16	P17		P19			P22
	B9		P2		P4		P6	P7		P9													
Chr. 11, 17.00 Mb	B1	P1									P10		P12						P18			P21	P22
	B3			P3								P11		P13					P18		P20		
	B10	P1				P5			P8		P10												
	B12	P1	P2				P6						P12										
Chr. 14, 11.13 Mb	B1	P1									P10		P12						P18			P21	P22
	B10	P1				P5			P8		P10												
Chr. 17 , 11.13 Mb	B1	P1									P10		P12						P18			P21	P22
	B12	P1	P2				P6						P12										
Chr. 18, 17.01 Mb	B4				P4					P9		P11			P14	P15				P19			
	B11			P3	P4							P11											
Chr. 22 , 4.97 Mb	B10	P1				P5			P8		P10												
	B12	P1	P2				P6						P12										
Chr. 22, 8.07 Mb	B4				P4					P9		P11			P14	P15				P19			
	B9		P2		P4		P6	P7		P9													
	B11			P3	P4							P11											
Chr. 23, 13.44 Mb	B4				P4					P9		P11			P14	P15				P19			
	B9		P2		P4		P6	P7		P9													

FIGURE 6
Chromosomes and sites with more than one QTL for resistance to Psa in common between analyses. The parents contributing alleles to each QTL can be determined for some peaks by analysing the families that contribute to each bulk. The QTL, highlighted in green, from Chromosome 4 was likely from the parent P1. But the indication of parent P1 being the main contributor to QTL peaks on Chromosome 14 at 11.13 Mb and Chromosome 22 at 4.97 Mb may be misleading as the bulk B10 had a low contribution from parent P1. This may not have had a strong enough signal to present as a peak unless the loci were shared with another parent such as P12. The QTL on Chromosome 2 at 5.35 Mb was likely from parent P12, and the peak on Chromosome 22 at 8.07 Mb was likely from parent P4. Other parents contributing to QTLs in light green had two individuals that could have contributed to the QTL. Chromosome 11 had no parental contributors to bulks in common with the QTL at that position, despite four bulks presenting QTL at that site. It is possible that the parents, P1 and P18, contained the same alleles for Psa resistance on Chromosome 11 at 16.95 Mb.

10.3389/fpls.2023.1255506

13.02 Mb. The parent P1 also appears to be the contributing parent to the QTL on Chromosome 22 at 4.97 Mb. However, this prediction of the P1 parent contribution to QTL at 4.97 Mb on Chromosome 22 may be inaccurate as the P1 parent makes up only a small proportion of the bulk B10 (5.6%). Instead, both the parents P6 and P8 may have contributed this QTL to these bulks. Parent P8 was used as one of the parents in a biparental mapping family for Psa resistance in a study by Tahir et al. (2019) and parents P6 and P8 are related. Tahir et al.'s study (2019) also identified the same QTL at 4.97 Mb on Chromosome 22 derived from parent P8. Eight other QTLs could have their parent contributors narrowed down to only two parents since both were shared between bulks and sites (Figure 6).

In cases where a QTL was detected in only one bulk, the allele responsible may have been contributed by a parent unique to that bulk (Figure 7).

Therefore, the QTL peaks on Chromosomes 11 at 0.36 Kb, 13 at 11.55 Mb, 26 at 9.12 Mb, 27 at 8.46 Mb and 29 at 10.68 Mb from bulk B9 were likely contributed by the parent P7. The remaining QTL sites from bulks B2, B3, B4 and B10 each had two unique parents that likely contributed to the detected QTL: namely, parents P16 or P17 contributed to Chromosomes 3 at 18.56 Mb, 6 at 12.65 Mb, 8 at 12.07 Mb, 19 at 23.56 Mb, 21 at 7.45 Kb, and 27 at 4.76 Kb, parents P13 or P17 to Chromosomes 11 at 7.97 Mb and 12 at 13.65 Mb, parents P13 or P20 to Chromosomes 7 at 16.92 Mb and 13 at 13.65 Mb from bulk B3, and parents P5 or P8 likely contributed to resistance alleles on Chromosomes 5 at 10.95 Mb, 17 at 13.31 Mb and 24 at 9.12 Mb from bulk B10. It was reassuring that the B11 and B12 bulks, which had no parents unique to the bulk, had no unique resistance allele sites attributed to them.

Further information about parent contributors to resistance alleles can be gained by identifying the parents in common among bulks that contributed to these alleles. This approach identified the likely parental contributors to three resistance alleles on Chromosomes 2, 4, and 22 from parents P1, P4 and P12, respectively (Table 3).

Effects of the selective sweep for Psa tolerance alleles

Within each sample bulk, the families that had more individuals surviving Psa contributed more DNA to the bulk compared to those with fewer surviving individuals included in the same bulk. When performing the BSA, bulks with a skewed family representation may have preferentially identified loci from the families with more individuals in the bulk. This is likely because a higher amount of DNA contributed to a site from a particular parent increases the read depth of a locus unique to that parent compared to the bulk of parents. This is what was expected for the resistant alleles, but the families that had fewer surviving individuals would be under-represented in the bulk and therefore the Gprime value may be lower for these loci. The lower Gprime value may be excluded at a locus of interest due to families with greater representation presenting higher Gprime values over a greater number of loci.

The individuals that contributed to bulks all survived the selective sweep caused by Psa. The selective sweep would have exerted strong selection for alleles linked to resistance loci, such as those at 16.95 Mb along Chromosome 11 in bulks 1, 3, 10, and 12. Conversely, the selective sweep would have significantly reduced the frequency of alternative alleles at those loci (Nielsen et al., 2005). With strong selection pressure for an allele from a parent contributing to the family, the other allele would be effectively eliminated from the population at that locus. However, changes in allele frequency can also be indirectly caused through genetic correlations from linkage disequilibrium (Barrett and Hoekstra, 2011; Kemppainen et al., 2017) and genetic drift (Conolly et al., 2008). The Gprime method of BSA was implemented to adjust for the effects of linkage disequilibrium (Magwene et al., 2011), but genetic drift could have skewed the results, particularly in families with poor representation in the bulk. This is because the families with poor representation in the bulk are also a poor representation of that family, which will predispose the alleles from these families to genetic drift (Magwene et al., 2011). Similar effects will have occurred at genomic regions linked to the alleles for Psa resistance (Robertson, 1970). However, because the effects of genetic drift are assumed to be random throughout the genome (Conolly et al., 2008), the effect of genetic drift on the results from this work are assumed to be minor and resistance alleles detected in multiple bulks are unlikely to be caused by genetic drift.

Identifying the parents that contributed the Psa resistance loci to each bulk will help with family-based breeding strategies (Hinds et al., 2005). Although the families that contributed the largest number of individuals to bulks are likely to be those that are contributing the alleles for resistance in each bulk, these resistance alleles could be coming from one or many parents. An attempt was made to identify the parents contributing resistance to the sample bulks analysed, but a combination of the missing parental WGS data and the way the bulks were constructed limited the information available. This could be overcome by developing markers to target the loci with high Gprime values. The markers could then be used on DNA from parents to identify

Bulk name	Parents contributing to bulks																				
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20	P21	P22
B1	P1									P10		P12						P18			P21	P22
B2		P2							P9			P12				P16	P17		P19			P22
B3			P3								P11		P13					P18		P20		
B4				P4					P9		P11			P14	P15				P19			
B9		P2		P4		P6	P7		P9													
B10	P1				P5			P8		P10												
B11			P3	P4							P11											
B12	P1	P2				P6						P12										

FIGURE 7
Bulks with parent contributors unique to each bulk. Parents that were represented in a single pool are highlighted in green.

TABLE 3 The commonality of parents contributing resistance loci among bulks.

Parent contributing to bulks	P1	P4	P12
Bulks containing parents	B1	B4	B1
	B10	B9	B2
	B12	B11	B12
QTL position	Chr. 4, 13.02 Mb	Chr. 22, 8.07 Mb	Chr. 2, 5.35 Mb

The parents that contributed specific alleles can be inferred where multiple bulks have resistance alleles at the same site that shared parents among those bulks. Parent P1 was the likely contributor to the resistance allele on Chromosome 4. Parent P4 was the likely contributor to the resistance allele on Chromosome 22, and Parent 12 was the likely contributor to the resistance allele on Chromosome 2.

which parents contributed these resistance alleles and enable marker-assisted selection for Psa resistance in families related to these parents. Identifying parents that contributed causal resistance alleles to a bulk could also be done by identifying alleles of higher frequency from the BSA under the QTL that were private to a single parent (Hinds et al., 2005). However, this was not possible in this work because many of the parents did not have WGS data available (Figure 2). Identifying haplotypes for each parent would also be beneficial by allowing the haplotype sequence of each parent to be matched with loci with high Gprime values under QTL. This would be informative for the parents of bulks B1, B4 and B11, but WGS data from parents P5, P6, P7, P16, P18, P19 and P22 would still be needed to generate haplotype sequences to identify the parents contributing to QTL in bulks B2, B3, B9, B10, and B12.

QTL detection accuracy of loci from parents that contributed a small or large percentage of DNA to bulks

Accurately detecting loci for resistance to Psa in part two of this work was dependent on the percentage of alleles that each family contributed to the bulk. Families that contributed more than 20% to a bulk and had strong allelic selection pressure on alleles unique to that family are likely to have those alleles present as a QTL in 72% of analyses (Figure 5). But, families that had strong selection pressure on alleles unique to that family and contributed 10–20% to a bulk were likely to present as a QTL in only 40% of analyses. Families that contributed less than 10% to a bulk were unlikely to present any QTL in the BSA, even with strong selection pressure on alleles unique to that family (Figure 5). Therefore, it is unlikely that families with poor representation in the sample bulk contributed to QTL in part two of this work. However, these families with low contribution to bulks may contribute the same resistance loci as other parents included in the bulk, adding to the significance of those sites. The lack of representation from the families contributing less than 10% to a bulk was due to the dilution created by other families that made up a bulk. For example, if an individual contributed 10% to a bulk and 50% of alleles were from one heterozygous parent under strong selection for Psa resistance, the resistance allele might be in all the individuals sampled from that family. Conversely, where resistance alleles are shared between families contributing to a bulk, their contribution to the sample bulk would stack, making their representation in the sample bulk 50% higher than in the parental bulk.

Future research

Future projects could improve upon the methods used in part one of this work. Sampling of individuals for the bulks containing different sex ratios should have been done on families where parental WGS data were available for all of the parents contributing to the families used in each bulk. If this were done when sampling in part one this work, the sample bulks with a differing number of males and females in each bulk could have been compared against the bioinformatically created bulk of parents to match part two of this work. However, this was not done because some of the parents that contributed to these bulks did not have WGS data. Having an accurate test of the methods would alleviate the concern that the inferences made in part one of this experiment reflect only the allelic variance included in the sampling and DNA extraction methodology and may not accurately reflect the influence of the BSA methods on identifying alleles for Psa resistance in part two of this work. If part two of this work were repeated, collecting leaf material from plants before being culled would create a better match to a typical BSA (Li and Xu, 2022) and allow the creation of a bulk of alleles that were being selected against instead of using a pool of parental WGS data as the comparison bulk. Also, collecting leaf samples from all plants before they were affected by Psa would enable the creation of an unselected bulk, a negatively selected bulk, and a positively selected bulk. Comparing the positively selected bulk of individuals resistant to Psa against the unselected bulk may allow the identification of alleles associated with resistance to Psa. Comparing the negatively selected bulk of individuals that were removed because of Psa against the unselected bulk may allow the identification of alleles associated with susceptibility to Psa. Integrating the crossing and sampling plans would increase the accuracy of analyses because the unselected bulk would be a better representation of the alleles within the family than that of bulks of WGS data from parents.

The two parts of this work showed that finding multiple alleles for resistance to Psa can be achieved using BSA of bulks containing multiple families while greatly simplifying field sampling, DNA extraction, and reducing sequencing costs. To our knowledge, this is the first time DNA has been extracted as a bulk for a BSA instead of quantifying DNA from each individual separately and bulking the resulting DNA. This is also the first time BSA has been applied to bulks of families, where the comparison bulk was made up of a bioinformatically generated bulk of parental WGS data to identify loci affecting the target trait. Utilising the alleles for Psa resistance found in this work as selection criteria in breeding programmes may enable faster breeding of cultivars with greater resistance to Psa than without marker-assisted selection and provide an opportunity to stack resistance loci to create a more robust resistance to Psa in future cultivars.

Data availability statement

The data presented in the study are deposited in the http://www.ncbi.nlm.nih.gov repository, accession number PRJNA1077747.

Author contributions

CF: Conceptualization, Data curation, Formal Analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. VS: Supervision, Writing – review & editing. RS: Data curation, Writing – review & editing. MD: Writing – review & editing, Data curation, Visualization. PD: 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 a PhD grant supplied to the corresponding author by The New Zealand Institute for Plant and Food Research, New Zealand.

Conflict of interest

Authors CF, RS and PD were employed by The New Zealand Institute for Plant and Food Research Limited.

The remaining author declares 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.2023.1255506/full#supplementary-material

References

Akagi, T., Henry, I. M., Ohtani, H., Morimoto, T., Beppu, K., Kataoka, I., et al. (2018). A Y-encoded suppressor of feminization arose via lineage-specific duplication of a cytokinin response regulator in kiwifruit. *Plant Cell* 30, 780–795. doi: 10.1105/tpc.17.00787

Barrett, R. D., and Hoekstra, H. E. (2011). Molecular spandrels: tests of adaptation at the genetic level. *Nat. Rev. Genet.* 12, 767. doi: 10.1038/nrg3015

Chen, N., Juric, I., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., Schoech, S. J., et al. (2019). Allele frequency dynamics in a pedigreed natural population. *Proc. Natl. Acad. Sci.* 116, 2158–2164. doi: 10.1073/pnas.1813852116

Conolly, J., Colledge, S., and Shennan, S. (2008). Founder effect, drift, and adaptive change in domestic crop use in early Neolithic Europe. *J. Archaeological Sci.* 35, 2797–2804. doi: 10.1016/j.jas.2008.05.006

Dai, P., Kong, J., Wang, S., Lu, X., Luo, K., Cao, B., et al. (2018). Identification of SNPs associated with residual feed intake from the muscle of *Litopenaeus vannamei* using bulk segregant RNA-seq. *Aquaculture* 497, 56–63. doi: 10.1016/j.aquaculture.2018.07.045

Datson, P., Nardozza, S., Manako, K., Herrick, J., Martinez-Sanchez, M., Curtis, C., et al. (2015). Monitoring the Actinidia germplasm for resistance to Pseudomonas syringae pv. actinidiae. *Acta Hortic.* 1095, 181–184. doi: 10.17660/ActaHortic.2015.1095.22

Deokar, A., Sagi, M., Daba, K., and Tar'an, B. (2019). QTL sequencing strategy to map genomic regions associated with resistance to ascochyta blight in chickpea. *Plant Biotechnol. J.* 17, 275–288. doi: 10.1111/pbi.12964

Dong, W., Wu, D., Li, G., Wu, D., and Wang, Z. (2018). Next-generation sequencing from bulked segregant analysis identifies a dwarfism gene in watermelon. *Sci. Rep.* 8, 1–7. doi: 10.1038/s41598-018-21293-1

Dwiartama, A. (2017). Resilience and transformation of the New Zealand kiwifruit industry in the face of Psa-V disease. *J. Rural Stud.* 52, 118–126. doi: 10.1016/j.jrurstud.2017.03.002

Everett, K. R., Taylor, R. K., Romberg, M. K., Rees-George, J., Fullerton, R. A., Vanneste, J. L., et al. (2011). First report of *Pseudomonas syringae* pv. *actinidiae* causing kiwifruit bacterial canker in New Zealand. *Australas. Plant Dis. Notes* 6, 67–71. doi: 10.1007/s13314-011-0023-9

Guan, L., Fan, P., Li, S.-H., Liang, Z., and Wu, B.-H. (2019). Inheritance patterns of anthocyanins in berry skin and flesh of the interspecific population derived from teinturier grape. *Euphytica* 215, 1–14. doi: 10.1007/s10681-019-2342-4

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. doi: 10.1016/bs.adgen.2018.12.001

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., et al. (2005). Whole-genome patterns of common DNA variation in three human populations. *Science* 307, 1072–1079. doi: 10.1126/science.1105436

James, J. (1970). The founder effect and response to artificial selection. *Genet. Res.* 16, 241–250. doi: 10.1017/S0016672300002500

Jansen, R. C. (1996). Complex plant traits: time for polygenic analysis. *Trends Plant Sci.* 1, 89–94. doi: 10.1016/S1360-1385(96)80040-9

Johnsson, M. (2018). Integrating selection mapping with genetic mapping and functional genomics. *Front. Genet.* 9, 603. doi: 10.3389/fgene.2018.00603

Kemppainen, P., Rønning, B., Kvalnes, T., Hagen, I. J., Ringsby, T. H., Billing, A. M., et al. (2017). Controlling for P-value inflation in allele frequency change in experimental evolution and artificial selection experiments. *Mol. Ecol. Resour.* 17, 770–782. doi: 10.1111/1755-0998.12631

Koh, Y., Chung, H., Cha, B., and Lee, D. (1994). Outbreak and spread of bacterial canker in kiwifruit. *Korean J. Plant Pathol. (Korea Republic)* 10, 68–72.

Konishi, T., Matsukuma, S., Fuji, H., Nakamura, D., Satou, N., and Okano, K. (2019). Principal component analysis applied directly to sequence matrix. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-55253-0

Lefebvre, V., and Palloix, A. (1996). Both epistatic and additive effects of QTLs are involved in polygenic induced resistance to disease: a case study, the interaction pepper—*Phytophthora capsici* Leonian. *Theor. Appl. Genet.* 93, 503–511. doi: 10.1007/BF00417941

Li, Z., and Xu, Y. (2022). Bulk segregation analysis in the NGS era: a review of its teenage years. *Plant J.* 109, 1355–1374. doi: 10.1111/tpj.15646

Magwene, P. M., Willis, J. H., and Kelly, J. K. (2011). The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput. Biol.* 7, e1002255. doi: 10.1371/journal.pcbi.1002255

Mansfeld, B. N., and Grumet, R. (2018). QTLseqr: an R package for bulk segregant analysis with next-generation sequencing. *Plant Genome* 11, 180006. doi: 10.3835/plantgenome2018.01.0006

Marsal, G., Boronat, N., Canals, J. M., Zamora, F., and Fort, F. (2013). Comparison of the efficiency of some of the most usual DNA extraction methods for woody plants in different tissues of *Vitis vinifera* L. *OENO One* 47, 227–237. doi: 10.20870/oeno-one.2013.47.4.1559

Matsumoto, Y., Goto, T., Nishino, J., Nakaoka, H., Tanave, A., Takano-Shimizu, T., et al. (2017). Selective breeding and selection mapping using a novel wild-derived heterogeneous stock of mice revealed two closely-linked loci for tameness. *Sci. Rep.* 7, 4607. doi: 10.1038/s41598-017-04869-1

McCann, H. C., Rikkerink, E. H., Bertels, F., Fiers, M., Lu, A., Rees-George, J., et al. (2013). Genomic analysis of the kiwifruit pathogen *Pseudomonas syringae* pv. *actinidiae* provides insight into the origins of an emergent plant disease. *PloS Pathog.* 9, e1003503. doi: 10.1371/journal.ppat.1003503

Michelmore, R. W., Paran, I., and Kesseli, R. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci.* 88, 9828–9832. doi: 10.1073/pnas.88.21.9828

Miko, I. (2008). Gregor Mendel and the principles of inheritance. *Nat. Educ.* 1, 134.

Munjal, G., Hao, J., Teuber, L. R., and Brummer, E. C. (2018). Selection mapping identifies loci underpinning autumn dormancy in alfalfa (*Medicago sativa*). *G3: Genes Genomes Genet.* 8, 461–468. doi: 10.1534/g3.117.300099

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575. doi: 10.1101/gr.4252305

Robertson, A. (1970). "A theory of limits in artificial selection with many linked loci," in Ki. Kojima. (eds) *Mathematical Topics in Population Genetics*. Biomathematics, Springer, Berlin, Heidelberg, 1, 246–288. doi: 10.1007/978-3-642-46244-3_8

Shen, R., and Messer, P. W. (2022). Predicting the genomic resolution of bulk segregant analysis. *G3* 12, jkac012. doi: 10.1093/g3journal/jkac012

Song, J., Li, Z., Liu, Z., Guo, Y., and Qiu, L.-J. (2017). Next-generation sequencing from bulked-segregant analysis accelerates the simultaneous identification of two qualitative genes in soybean. *Front. Plant Sci.* 8, 919. doi: 10.3389/fpls.2017.00919

Soto-Cerda, B. J., and Cloutier, S. (2012). "Association mapping in plant genomes," in *Genetic diversity in plants* (InTech).

Sun, J., Yang, L., Wang, J., Liu, H., Zheng, H., Xie, D., et al. (2018). Identification of a cold-tolerant locus in rice (*Oryza sativa* L.) using bulked segregant analysis with a next-generation sequencing strategy. *Rice* 11, 1–12. doi: 10.1186/s12284-018-0218-1

Tahir, J., Brendolise, C., Hoyte, S., Lucas, M., Thomson, S., Hoeata, K., et al. (2020). Qtl mapping for resistance to cankers induced by *Pseudomonas syringae* pv. *actinidiae* (psa) in a tetraploid *Actinidia chinensis* kiwifruit population. *Pathogens* 9, 967. doi: pathogens9110967

Tahir, J., Gardiner, S. E., Bassett, H., Chagné, D., Deng, C. H., and Gea, L. (2018). Tolerance to Pseudomonas syringae pv. actinidiae in a kiwifruit breeding parent is conferred by multiple loci. *Acta Hortic.* 1203, 67–70. doi: 10.17660/ActaHortic.2018.1203.10

Tahir, J., Hoyte, S., Bassett, H., Brendolise, C., Chatterjee, A., Templeton, K., et al. (2019). Multiple quantitative trait loci contribute to resistance to bacterial canker incited by *Pseudomonas syringae* pv. *actinidiae* in kiwifruit (*Actinidia chinensis*). *Horticulture Res.* 6, 1–18. doi: 10.1038/s41438-019-0184-9

Takikawa, Y., Serizawa, S., Ichikawa, T., Tsuyumu, S., and Goto, M. (1989). *Pseudomonas syringae* pv. *actinidiae* pv. nov. the causal bacterium of canker of kiwifruit in Japan. *Japanese J. Phytopathol.* 55, 437–444. doi: 10.3186/jjphytopath.55.437

Wang, Z., Yu, A., Li, F., Xu, W., Han, B., Cheng, X., et al. (2021). Bulked segregant analysis reveals candidate genes responsible for dwarf formation in woody oilseed crop castor bean. *Sci. Rep.* 11, 1–15. doi: 10.1038/s41598-021-85644-1

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: Indian J. Statistics Ser. A* Indian Statistical Institute 26, 359–372.

Wisser, R. J., Murray, S. C., Kolkman, J. M., Ceballos, H., and Nelson, R. J. (2008). Selection mapping of loci for quantitative disease resistance in a diverse maize population. *Genetics* 180, 583–599. doi: 10.1534/genetics.108.090118

Xin, F., Zhu, T., Wei, S., Han, Y., Zhao, Y., Zhang, D., et al. (2020). QTL mapping of kernel traits and validation of a major QTL for kernel length-width ratio using SNP and bulked segregant analysis in wheat. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-019-56979-7

Yao, Z., You, F. M., N'diaye, A., Knox, R. E., Mccartney, C., Hiebert, C. W., et al. (2020). Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinf.* 21, 1–16. doi: 10.1186/s12859-020-03704-1

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 10 April 2024
DOI 10.3389/fpls.2024.1320844

OPEN ACCESS

EDITED BY
Satoshi Watanabe,
Saga University, Japan

REVIEWED BY
Gezahegn Girma,
Purdue University, United States
Yuri Shavrukov,
Flinders University, Australia
Jindong Liu,
Chinese Academy of Agricultural Sciences, China

*CORRESPONDENCE
Yi-Hong Wang
yihong.wang@louisiana.edu
Jieqin Li
wlhljq@163.com

RECEIVED 13 October 2023
ACCEPTED 20 March 2024
PUBLISHED 10 April 2024

CITATION
Wang L, Tu W, Jin P, Liu Y, Du J, Zheng J, Wang Y-H and Li J (2024) Genome-wide association study of plant color in *Sorghum bicolor*. *Front. Plant Sci.* 15:1320844. doi: 10.3389/fpls.2024.1320844

Genome-wide association study of plant color in *Sorghum bicolor*

Lihua Wang[1,2], Wenmiao Tu[1,2], Peng Jin[1,2], Yanlong Liu[1,2], Junli Du[1,2], Jiacheng Zheng[1,2], Yi-Hong Wang[3]* and Jieqin Li[1,2]*

[1]College of Agriculture, Anhui Science and Technology University, Fengyang, Anhui, China, [2]Anhui Province International Joint Research Center of Forage Bio-breeding, Chuzhou, China, [3]Department of Biology, University of Louisiana at Lafayette, Lafayette, LA, United States

Introduction: Sorghum plant color is the leaf sheath/leaf color and is associated with seed color, tannin and phenol content, head blight disease incidence, and phytoalexin production.

Results: In this study, we evaluated plant color of the sorghum mini core collection by scoring leaf sheath/leaf color at maturity as tan, red, or purple across three testing environments and performed genome-wide association mapping (GWAS) with 6,094,317 SNPs markers.

Results and Discussion: Eight loci, one each on chromosomes 1, 2, 4, and 6 and two on chromosomes 5 and 9, were mapped. All loci contained one to three candidate genes. In *qPC5-1*, Sobic.005G165632 and Sobic.005G165700 were located in the same linkage disequilibrium (LD) block. In *qPC6*, Sobic.006G149650 and Sobic.006G149700 were located in the different LD block. The single peak in *qPC6* covered one gene, Sobic.006G149700, which was a senescence regulator. We found a loose correlation between the degree of linkage and tissue/organ expression of the underlying genes possibly related to the plant color phenotype. Allele analysis indicated that none of the linked SNPs can differentiate between red and purple accessions whereas all linked SNPs can differentiate tan from red/purple accessions. The candidate genes and SNP markers may facilitate the elucidation of plant color development as well as molecular plant breeding.

KEYWORDS

GWAS, plant color, resequencing, sorghum, SNP

1 Introduction

Plant color in sorghum [*Sorghum bicolor* (L.) Moench] is defined as the stem/leaf sheath/leaf color (Rana et al., 1976; Reddy et al., 2008; Rooney, 2016; Fedenia et al., 2020) at maturity (Valencia and Rooney, 2009). Plant color is controlled by the *P* and *Q* genes. A

sorghum plant with *P_Q_* genotype is purple, whereas *P_ qq* is red and *pp Q_* and *pp qq* are tan (Dykes et al., 2009; Valencia and Rooney, 2009; Dykes et al., 2011).

Plant color is associated with other phenotypes or consumer preferences. For example, white sorghum grain from tan plants is more desirable for human or animal consumption (Williams-Alanis et al., 1999; Funnell and Pedersen, 2006; Rooney, 2016). This is probably because tan plants tend to have lower tannin content compared with purple plants (Gourley and Lusk, 1978; Dykes et al., 2005). However, sorghum grains grown on plants with purple/red plant color do have higher levels of total phenols than those from tan plants (Dykes et al., 2005), although grains from some tan plants have the highest flavone (luteolin and apigenin) content (Dykes et al., 2009, Dykes et al., 2011). Tan plants tend to have lower head blight incidence caused by *Fusarium moniliforme* than red plants (Torres-Montalvo et al., 1992), but it is not clear if this is related to the high luteolin and apigenin contents. Du et al. (2010) have shown that flavones such as luteolin function as a phytoalexin against the sorghum anthracnose pathogen *Colletotrichum sublineolum*.

Sorghums with red/purple plant color produce the highest levels of 3-deoxyanthocyanidins (apigeninidin and luteolinidin) (Dykes et al., 2011), which are also phytoalexins induced by fungal attack (Snyder and Nicholson, 1990). The purple phenotype after fungal attack is determined by the production of two 3-deoxyanthocyanidins, apigeninidin and luteolinidin, which are not produced by the tan plants (Kawahigashi et al., 2016). The underlying *P* gene has been cloned using map-based cloning in progeny from a cross between purple Nakei-MS3B (*PP*) and tan Greenleaf (*pp*) cultivars; the gene was located in a 27-kb genomic region between markers CA29530 and SB25792 on chromosome 6. Four candidate genes identified in this region were similar to the maize leucoanthocyanidin reductase gene induced by wounding, and only the Sb06g029550 gene was induced in both cultivars after wounding. The Sb06g029550 protein was detected in Nakei-MS3B but only slightly in Greenleaf. A recombinant Sb06g029550 protein had a specific flavanone 4-reductase activity and converted flavanones (naringenin or eriodictyol) to flavan-4-ols (apiforol or luteoforol) *in vitro* (Kawahigashi et al., 2016).

In this study, we evaluated plant color of the sorghum mini core collection (MC; Upadhyaya et al., 2009) as the association panel. This panel has been extensively characterized, such as its genetic structure and linkage disequilibrium (Wang et al., 2013) and effectiveness for association mapping (Upadhyaya et al., 2013). Most importantly, the panel has been used to clone a pleiotropic *SbSNF4-2* (*SnRK1βγ2*) that increases both biomass and sugar yield in sorghum and sugarcane (Upadhyaya et al., 2022). We scored leaf sheath/leaf color at maturity as tan, red, or purple across three testing environments in Tengqiao/Hainan and Fengyang/Anhui, China, performed association mapping with 6,094,317 SNP markers (Wang et al., 2021), and identified candidate genes strongly linked to plant color.

2 Materials and methods

2.1 Plant materials and phenotyping

The accessions of the sorghum MC (Upadhyaya et al., 2009, Table S1) were grown in Tengqiao, Hainan, China, for two seasons (2021 and 2022) and in Fengyang, Anhui, China, for one season (2022). In both 2021 and 2022 in Tengqiao, Hainan, the plants were grown with a row spacing of 65 cm and a plant spacing within each row of 25 cm. A compound fertilizer (N:P:K = 15:15:15) and urea were applied before planting at 200 kg/ha and 120 kg/ha, respectively. The plot was irrigated once at seedling and once at stem elongation stages and weeded at before three-leaf, during four-to-six-leaf, and before anthesis stages. Pesticides were applied three times to control cutworms, aphids, and honeydew moths.

In Fengyang, Anhui in 2022, the plants were grown with a row spacing of 50 cm and plant spacing within each row of 25 cm. A compound fertilizer (N:P:K = 15:15:15) and urea were applied before planting at 180 kg/ha and 90 kg/ha, respectively. The plot was irrigated once at seedling and once at stem elongation stages and weeded at before three-leaf, during four-to-six-leaf, and before anthesis stages. Pesticides were applied three times to control cutworms, aphids, and honeydew moths.

FIGURE 1
The sorghum plant color phenotype from the mini core collection: tan, red, and purple according to Rooney (2016).

At maturity in all three environments, plant color was scored for leaf/leaf sheath color as "1" (tan), "2" (red), or "3" (purple) (Figure 1) according to Rooney (2016).

2.2 Genome-wide association study

Genome resequencing of 237 MC accessions (Supplementary Table S1) and genome-wide association study (GWAS) were as described in Wang et al. (2021). GWAS was performed with 6,094,317 SNPs from Wang et al. (2021). The kinship matrix (*K*) was generated by EMMAX (Kang et al., 2010), which was used to perform GWAS analyses with the Q matrix calculated using STRUCTURE 2.3.4 (Pritchard et al., 2000) as the covariate variable. The modified Bonferroni correction was used to determine the genome-wide significance thresholds of the GWAS, based on a nominal level of $\alpha = 0.05$ which corresponds to a *P* value of 8.2E-09, or $-\log_{10}(P)$ values of 8.08. At $\alpha = 0.01$, these were 1.6E-09 and 8.78, respectively.

2.3 Candidate gene identification and allelic effect of linked SNPs

Candidate genes were identified using the reference genome Sorghum bicolor v3.1.1 (Paterson et al., 2009; McCormick et al., 2018) curated at Phytozome (Goodstein et al., 2012) 13 (https://phytozome-next.jgi.doe.gov/). RNA-seq data (McCormick et al., 2018) for each candidate genes were downloaded from the site and provided as Supplementary Table S2. To determine the allelic effect of selected SNPs linked to plant color, SNPs in each locus or two loci were grouped together. Only accessions with less than 5% missing data rate for each group of SNPs were included. The original data are provided in Supplementary Tables S3-S8.

3 Results

3.1 Phenotype analysis

As described in the Introduction, plant color is controlled by multiple genes. This is reflected in phenotyping in this study. All accessions were consistently scored as either tan (9 accessions) or pigmented (228 accessions) in all three environments (2021_HN, 2022_HN, 2022_FY; Supplementary Table S1). However, 47 of the 228 accessions (20.6%) could not be consistently scored as either red or purple across the three environments. This indicates that the trait may be affected by the environment as well as the combinations of multiple genes.

3.2 Association mapping

To identify SNP markers linked to the trait, we used the following criteria: 1) more than one marker associated with plant color and at least one of the markers had $-\log_{10}(P)$ higher than the threshold (Upadhyaya et al., 2022), and 2) association had to be present across all three environments (2021_HN, 2022_HN, and 2022_FY). Based on these criteria, we identified eight loci distributed on chromosomes 1, 2, 4, 5, 6, and 9 (Figure 2; Supplementary Figure S1; Table 1). These loci contained 2 (*qPC2* and *qPC4*) to 21 SNP markers (*qPC5-2*) (Table 1). The strongest association was with the SNP (64621753) marker on chromosome 5 (*qPC5-2*), with $-\log_{10}(P)$ values of 11.50 in 2021_HN, 11.65 in 2022_HN, and 9.26 in 2022_FY (Table 1), respectively. This was followed by the locus on chromosome 6 for 51113980 (*PC6*) with $-\log_{10}(P)$ values of 10.4, 11.1, and 10.3, respectively (Table 1). *qPC5-1* and *qPC5-2* were mapped with the most SNPs with $-\log(P)$ values higher than 6.0, 20, and 21 SNPs (Table 1), respectively.

3.3 Candidate gene identification

Only genes closest to the respectively linked SNPs are presented in Table 1. All loci contained one to three candidate genes (Table 1). The *qPC5-1* and *qPC6* were further examined with linkage disequilibrium (LD) analysis combined with the Manhattan plot (Figure 3). In *qPC5-1*, Sobic.005G165632 and Sobic.005G165700 were located in the same LD block with the QTL peak. In *qPC6*, Sobic.006G149650 and Sobic.006G149700 were located in the different LD blocks. The *qPC6* peak contained only one gene, Sobic.006G149700, which indicates that it should be the candidate gene for *qPC6*. The annotation information showed that Sobic.006G149700 is senescence regulator/heavy metal-associated isoprenylated plant protein 34.

3.4 Allelic effect on plant color

We examined the allelic effect of all SNPs from the eight loci. For each locus, only accessions with missing data rate less than 5% were selected. In all loci, more purple accessions were observed than tan and red combined and no SNPs from the loci could differentiate between purple and red color accessions whereas most SNPs from all loci can differentiate tan from red/purple accessions (Supplementary Tables S3-S8). We presented three of four SNPs (5:64621753, 5:64224755, and 6:51113980) most tightly linked to plant color from Table 1 in Figure 4. Six tan accessions were identified for all three SNPs whereas 7, 12, and 5 red accessions were identified, respectively. In contrast, 37, 71, and 55 purple accessions were identified respectively for the three markers. In both 5:64224755 (T/C) and 6:51113980 (G/C), IS20740 was the single heterozygote and the T and G alleles respectively were dominant to the C alleles as CC homozygotes in both SNPs were red or purple, whereas the heterozygotes were tan. In the other five accessions, TT and GG genotypes in the two SNPs showed tan plant color. It is coincidental that in all three SNPs, red/purple accessions were all CC genotypes.

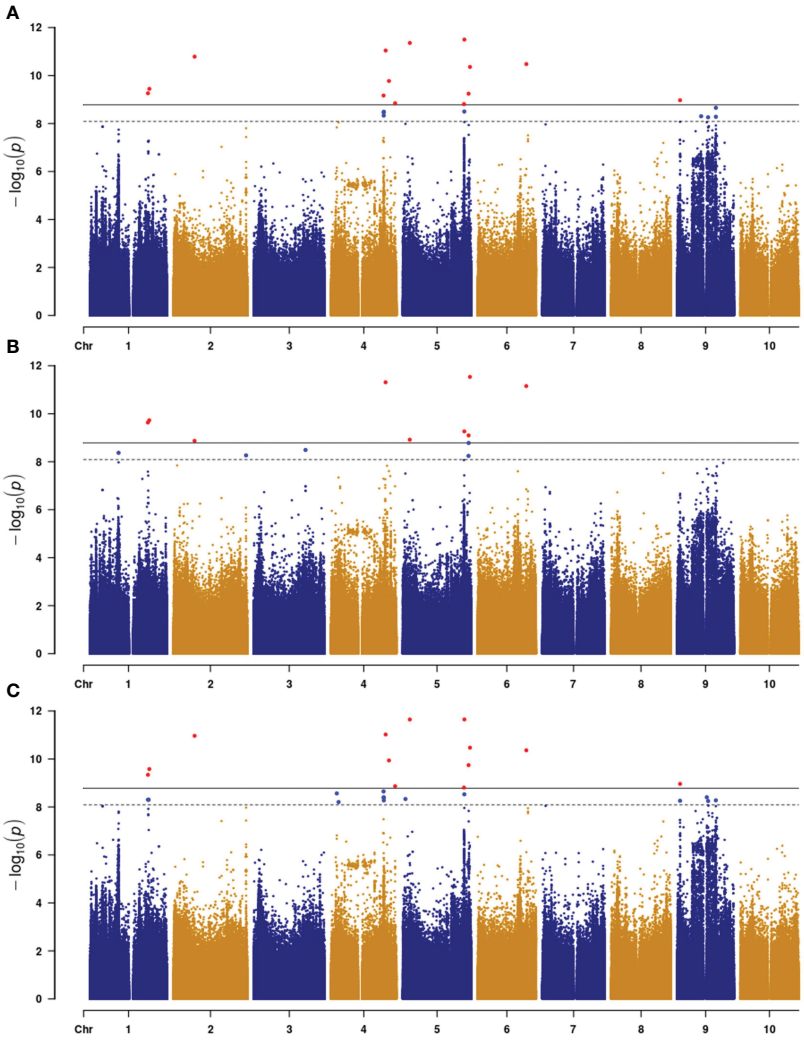


FIGURE 2
Manhattan plot of plant color with three environments in sorghum ((**A**), 2021_HN; (**B**), 2022_HN; (**C**), 2022_FY). Horizontal dash and gray lines indicate the threshold −log(*P*) value at α = 0.05 and 0.01, respectively.

4 Discussion

White sorghum grain grown on tan plants is highly desirable as livestock feed and for human consumption (Awika et al., 2002). The tan/purple/red plant color is mainly controlled by the *P* and *Q* genes (Dykes et al., 2009; Valencia and Rooney, 2009; Dykes et al., 2011). In this study, we identified eight loci for plant color across three environments. Among these, *qPC6* locus at 51,113,980 bp on chromosome 6 is long way off the plant color QTLs mapped by Boyles et al. (2017). They mapped one locus each at 56650607 and 56635333 bp on chromosome 6 in BTx642/BTxARG-1 and BTxARG-1/P850029 RIL populations, respectively. However, their two QTLs range from 49.9 Mb to 60.77 Mb and from 50.91 Mb to 60.6 Mb, both overlapping with *qPC6*. The peaks at 56,650,607 and 56,635,333 bp are close to the *P* gene (57164448.57187434 in *Sorghum bicolor* v3.1.1), which turns the leaves to purple upon wounding or pathogen invasion (Kawahigashi et al., 2016). This is because Boyles et al. (2017) used *Sorghum bicolor* v3.1 and Kawahigashi et al. (2016) used *Sorghum bicolor* v1.4 at www.plantgdb.org/SbGDB, which is no longer functional at the time of this writing. Therefore, genomic locations are not comparable although Sb06g029550 (Sobic.006G226800) from Kawahigashi et al. (2016) is located in *Sorghum bicolor* v3.1 as from 57,175,961 bp to 57,178,219 bp on chromosome 6. In *qPC6*, those highly associated SNPs were only located in the Sobic.006G149700 gene region (Figure 3), which is annotated as a senescence regulator. Its highest expression was in the leaf sheath at floral initiation, followed by seeds at maturity and juvenile leaf blades (Supplementary Table S2; McCormick et al., 2018). It is clear that *qPC6* does not overlap with the *P* gene. This could suggest that there are multiple genes responsible for plant color in sorghum. Sobic.006G149700 is orthologous to Arabidopsis *AtS40* (AT2G28400) and its mutation delayed leaf senescence (Fischer-Kilbienski et al., 2010).

10.3389/fpls.2024.1320844

TABLE 1 The plant color QTLs in sorghum detected in all three environments*.

SNPs	–log(*P*)			Candidate gene position	Candidate gene expression**
	HN 2021	HN 2022	FY 2022		
Chr01 (*qPC1*)					
61215127	6.84	7.92	6.51	Sobic.001G324900 DUF2215 Chr01:61207155.61212055 forward 3 kb from 61215127	Highest expression in leaves
61215130	7.25	**8.30**	6.28		
61215132	7.25	**8.30**	6.28	Sobic.001G325000 disease resistance protein Chr01:61220311.61222354 forward 5 kb from 61215155	Highly expressed in seeds and roots
61215133	7.25	**8.30**	6.28		
61215136	7.25	**8.30**	6.28		
61215145	6.82	7.69	6.79		
61215155	6.75	7.66	7.42		
Chr02 (*qPC2*)					
76356664	7.80	7.97	3.89	Sobic.002G416400 bHLH033 Chr02:76362451.76364029 forward Between the SNPs	Highest exp in internode and leaf sheath
76366175	7.44	7.43	**8.26**		
Chr04 (*qPC4*)					
55131019	**9.17**	**8.65**	7.29	Sobic.004G200700 ABI4 Chr04:55121061.55121994 forward 9 kb from 55131019	Highly expressed in panicles
55132948	7.27	7.49	3.01		
Chr05 (*qPC5-1*)					
64207536	7.19	6.90	6.16	Sobic.005G165632 unknown Chr05:64212726.64213572 reverse 64213028 in 5′-UTR and 64213268 in the intron	Highly expressed in panicles and seeds
64209111	7.10	6.79	6.09		
64209805	7.10	6.79	6.09		
64209962	7.06	6.76	6.07	Sobic.005G165700 Plant antimicrobial peptide (MBP-1 family protein precursor) Chr05:64214229.64216294 forward 64216240 in 3′-UTR	Panicle and seed specific expression
64210891	7.04	6.74	6.04		
64210945	7.33	7.05	6.09		
64211162	7.10	6.79	6.09		
64211239	7.10	6.79	6.09		
64211411	7.10	6.79	6.09		
64211531	7.10	6.79	6.09		
64213028	6.98	6.66	6.17	Sobic.005G165800 MSS1/GTP-binding protein Chr05:64217561.64225842 reverse 64224755 in second intron	Highly expressed in panicles and leaves
64213268	6.55	6.22	4.21		
64216240	7.02	6.97	6.16		
64216576	6.07	5.99	5.26		
64216992	6.91	6.83	6.07		
64217122	6.91	6.83	6.06		
64224755	**8.81**	**8.81**	8.07		
64266119	7.26	6.12	5.09		
64266282	7.26	6.12	5.09		
64267101	7.39	6.25	5.09		

(Continued)

TABLE 1 Continued

SNPs	–log(*P*)			Candidate gene position	Candidate gene expression**
	HN 2021	HN 2022	FY 2022		
Chr05 (*qPC5-2*)					
64580048	**8.50**	**8.53**	7.43	Sobic.005G167600 similar to Pi-b protein Chr05:64628186.64630007 reverse Between 64621753 and 64638422	Not highly expressed
64580306	6.72	6.60	5.45		
64581328	6.81	6.68	5.34		
64581344	6.81	6.68	5.34		
64583869	6.85	6.74	5.56		
64584322	8.06	7.96	7.00		
64584997	6.75	6.66	5.33		
64585014	6.73	6.64	5.33		
64586194	6.63	6.20	6.42		
64586498	7.38	6.84	5.93		
64587305	6.54	6.41	5.89		
64589960	6.96	6.87	5.66		
64590140	6.76	6.66	5.42		
64591940	6.68	6.56	5.44		
64610693	6.77	6.66	5.51		
64612690	6.93	6.86	5.51		
64612943	6.10	6.06	5.05		
64614467	5.44	5.29	5.37		
64616198	6.98	6.84	5.66		
64621753	**11.50**	**11.65**	**9.26**		
64638422	6.41	6.01	5.00		
Chr06 (*qPC6*)					
51113845	3.60	4.51	3.21	Sobic.006G149700 Senescence regulator Chr06:51115119.51116554 reverse Between 51113980 and 51116621	Highest expression in leaf sheath
51113980	**10.48**	**10.36**	**11.15**		
51114635	3.02	3.29	4.43		
51115418	1.79	2.08	4.45		
51115424	2.01	2.32	4.19		
51116621	3.28	4.26	3.09		
Chr09 (*qPC9-1*)					
2824326	**8.97**	**8.96**	**6.59**	Sobic.009G031700 unknown Chr09:2823951.2827282 reverse All three SNPs in coding region	Highly expressed in flowers and leaves
2824605	7.32	7.34	5.16		
2824643	6.43	6.36	5.51		
Chr09 (*qPC9-2*)					
40485401	**8.66**	**8.28**	7.10	Sobic.009G101700 RP-S7e Chr09:40497726.40501826 reverse 16.4 kb from 40485401	Ubiquitously expressed
40485626	7.59	7.20	6.21		
40486124	**8.28**	8.03	6.16		

(Continued)

10.3389/fpls.2024.1320844

TABLE 1 Continued

SNPs	–log(*P*)			Candidate gene position	Candidate gene expression**
	HN 2021	HN 2022	FY 2022		
Chr09 (*qPC9-2*)					
40487728	6.42	5.99	6.50		
40488163	7.39	7.00	5.83		
40488216	7.92	7.43	5.96		
40491677	6.36	5.91	6.39		
40491733	7.23	6.82	5.82		
40492084	6.29	6.20	4.60		
40492724	7.25	6.86	5.94		
40492742	7.25	6.86	5.94		

*SNP position is based on Sorghum bicolor v3.1.1. –log(P) in bold indicates significance at $\alpha = 0.05$ or 0.01. Underlined candidate genes are closest to the linked SNPs. ** See Supplementary Table S2 for expression data.

As mentioned above, RNA-seq expression data (Supplementary Table S2) by McCormick et al. (2018) may help identify candidate genes. In this study, plant color was scored for leaf/leaf sheath color as "1" (tan), "2" (red), or "3" (purple) according to Rooney (2016). Candidate genes physically close to the linked SNPs are either highly expressed in leaves, leaf sheath, or both (Table 1; Supplementary Table S2). For example, in *qPC1* Sobic.001G324900 is the only gene within 3 kb of the locus and the gene's highest expression is in the leaves and moderate expression in the leaf sheath; Sobic.002G416400 in *qPC2* is the only gene between the linked SNPs and is highly expressed in the leaf sheath; in *qPC5-1*, three genes are within the locus but only Sobic.005G165800 is highly expressed in both leaves and leaf sheath; as the only gene within the *qPC6* locus, Sobic.006G149700's highest expression is in the leaf sheath and leaves; and Sobic.009G031700 is the only gene in *qPC9-1* with all linked SNPs in its coding region and is highly expressed in the leaves. The only exception is Sobic.005G167600 in *qPC5-2*, which is the only gene within the linked SNPs, and it is not highly expressed. In contrast, Sobic.004G200700 is 9 kb from *qPC4* and is only highly expressed in the panicles and Sobic.009G101700 in *qPC9-2* is 16 kb away and ubiquitously expressed. These indicate a loose correlation between the degree of linkage and tissue/organ expression of the underlying genes. It is possible that altered expression of these genes could impact plant color scored using leaves and leaf sheath.

Sorghums with red/purple plant color are also induced by fungal attack (Snyder and Nicholson, 1990). In the current study, we also identified one candidate gene associated with fungal resistance. In *qPC5-2*, Sobic.005G165700 is the antimicrobial

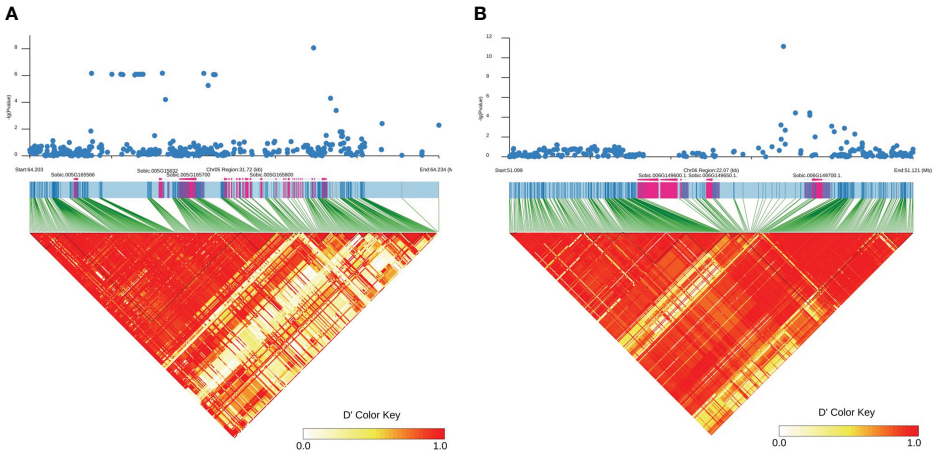


FIGURE 3
Manhattan plot (top) of the *qPC5-1* **(A)** and *qPC-6* **(B)** loci on chromosomes 5 and 6 and their respective LD blocks (bottom). The X-axis represents position in bp along the chromosome, and the Y-axis is $-\log_{10}(P)$.

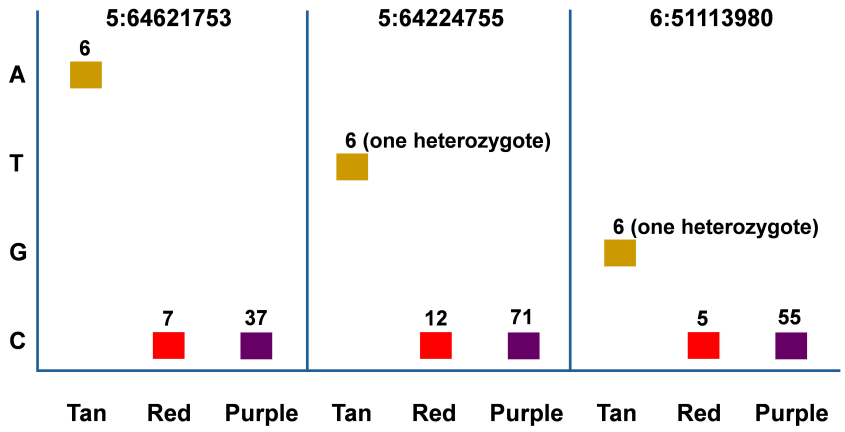


FIGURE 4
Allelic effect of SNP markers 5:64621753, 5:64224755, and 6:51113980 on sorghum plant color. The accessions with each plant color were selected to maximize the number of each color with minimum missing genotype data rate. Therefore, the accessions with the same color across different SNPs may overlap but may not be identical.

peptide MBP-1 family protein precursor, which has been reported as effective against both Gram-negative and Gram-positive bacteria as well as several filamentous fungi (Duvick et al., 1992). As stated above, Sobic.005G165800 is highly expressed in both leaves and leaf sheath, although its highest expression is in seed grain at maturity and the panicles (Supplementary Table S2). There is no ortholog of this gene in Arabidopsis, and no orthologs in maize or rice have been studied. Therefore, the correlation of plant color and antimicrobial peptide needs to be further investigated.

In conclusion, in this study, we mapped eight loci associated with sorghum plant color, one each on chromosomes 1, 2, 4, and 6 and two on chromosomes 5 and 9. We identified several candidate genes that are highly expressed in the leaves/leaf sheath, and one of the candidate genes was Sobic.006G149700 encoding a senescence regulator. This may facilitate the elucidation of plant color development as well as molecular plant breeding.

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

LW: Writing – review & editing. WT: Investigation, Writing – review & editing. PJ: Writing – review & editing. YL: Writing – review & editing. JD: Writing – review & editing. JZ: Writing – review & editing. Y-HW: Writing – original draft. JL: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that 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 (32372134), the Anhui Provincial Natural Science Fund (2008085MC73), the Key Project of Natural Science Research of Anhui Provincial Education Department (KJ2021ZD0108), and the Distinguished talents of Anhui Provincial Education Department (gxbjZD2022045).

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.2024.1320844/full#supplementary-material

References

Awika, J. M., Suhendro, E. L., and Rooney, L. W. (2002). Milling value of sorghums compared by adjusting yields to a constant product color. *Cereal Chem.* 79, 249–251. doi: 10.1094/CCHEM.2002.79.2.249

Boyles, R. E., Pfeiffer, B. K., Cooper, E. A., Zielinski, K. J., Myers, M. T., Rooney, W. L., et al. (2017). Quantitative trait loci mapping of agronomic and yield traits in two grain sorghum biparental families. *Crop Sci.* 57, 2443–2456. doi: 10.2135/cropsci2016.12.0988

Du, Y., Chu, H., Wang, M., Chu, I. K., and Lo, C. (2010). Identification of flavone phytoalexins and a pathogen-inducible flavone synthase II gene (*SbFNSII*) in sorghum. *J. Exp. Bot.* 61, 983–994. doi: 10.1093/jxb/erp364

Duvick, J. P., Rood, T., Rao, A. G., and Marshak, D. R. (1992). Purification and characterization of a novel antimicrobial peptide from maize (Zea mays L.) kernels. *J. Biol. Chem.* 267, 18814–18820. doi: 10.1016/S0021-9258(19)37034-6

Dykes, L., Peterson, G. C., Rooney, W. L., and Rooney, L. W. (2011). Flavonoid composition of lemon-yellow sorghum genotypes. *Food Chem.* 128, 173–179. doi: 10.1016/j.foodchem.2011.03.020

Dykes, L., Rooney, L. W., Waniska, R. D., and Rooney, W. L. (2005). Phenolic compounds and antioxidant activity of sorghum grains of varying genotypes. *J. Agric. Food Chem.* 53, 6813–6818. doi: 10.1021/jf050419e

Dykes, L., Seitz, L. M., Rooney, W. L., and Rooney, L. W. (2009). Flavonoid composition of red sorghum genotypes. *Food Chem.* 116, 313–317. doi: 10.1016/j.foodchem.2009.02.052

Fedenia, L., Klein, R. R., Dykes, L., Rooney, W. L., and Klein, P. E. (2020). Phenotypic, phytochemical, and transcriptomic analysis of black sorghum (*Sorghum bicolor* L.) pericarp in response to light quality. J. Agri. *Food Chem.* 68, 9917–9929. doi: 10.1021/acs.jafc.0c02657

Fischer-Kilbienski, I., Miao, Y., Roitsch, T., Zschiesche, W., Humbeck, K., and Krupinska, K. (2010). Nuclear targeted AtS40 modulates senescence associated gene expression in Arabidopsis thaliana during natural development and in darkness. *Plant Mol. Biol.* 73, 379–390. doi: 10.1007/s11103-010-9618-3

Funnell, D. L., and Pedersen, J. F. (2006). Association of plant color and pericarp color with colonization of grain by members of *Fusarium* and *Alternaria* in near-isogenic sorghum lines. *Plant Dis.* 90, 411–418. doi: 10.1094/PD-90-0411

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acid Res.* 40, D1178–D1186. doi: 10.1093/nar/gkr944

Gourley, L. M., and Lusk, J. W. (1978). Genetic parameters related to sorghum silage quality. *J. Dairy Sci.* 61, 1821–1827. doi: 10.3168/jds.S0022-0302(78)83808-9

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., 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

Kawahigashi, H., Kasuga, S., Sawada, Y., Yonemaru, J. I., Ando, T., Kanamori, H., et al. (2016). The sorghum gene for leaf color changes upon wounding (*P*) encodes a flavanone 4-reductase in the 3-deoxyanthocyanidin biosynthesis pathway. *G3 (Bethesda)* 6, 1439–1447. doi: 10.1534/g3.115.026104

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93, 338–354. doi: 10.1111/tpj.13781

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457, 551–556. doi: 10.1038/nature07723

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155, 945–959. doi: 10.1093/genetics/155.2.945

Rana, B. S., Tripathi, D. P., and Rao, N. G. P. (1976). Genetic analysis of some exotic x Indian crosses in sorghum. XV. Inheritance of resistance to sorghum rust. *Indian J. Genet. Plant Breed.* 36, 244–249.

Reddy, R. N., Mohan, S. M., Madhusudhana, R., Umakanth, A. V., Satish, K., and Srinivas, G. (2008). Inheritance of morphological characters in sorghum. *J. SAT Agric. Res.* 6, 1–3.

Rooney, W. L. (2016). "Sorghum: production and improvement practices," in *The Production and Genetics of Food Grains/Encyclopedia of Food Grains*, *2nd ed*, vol. 4, 2016. Academic Press, Waltham MA.

Snyder, B. A., and Nicholson, R. L. (1990). Synthesis of phytoalexins in sorghum as a site-specific response to fungal ingress. *Science* 248, 1637–1639. doi: 10.1126/science.248.4963.1637

Torres-Montalvo, H., Mendoza-Onofre, L., Gonzalez-Hernandez, V., and Williams-Alanis, H. (1992). Reaction of tan and non-tan isogenic genotypes to head blight. *Intl Sorghum Millets Newsl* 33, 36.

Upadhyaya, H. D., Pundir, R. P. S., Dwivedi, S. L., Gowda, C. L. L., Reddy, V. G., and Singh, S. (2009). Developing a mini core collection of sorghum for diversified utilization of germplasm. *Crop Sci.* 49, 1769–1780. doi: 10.2135/cropsci2009.01.0014

Upadhyaya, H. D., Wang, Y. H., Gowda, C. L., and Sharma, S. (2013). Association mapping of maturity and plant height using SNP markers with the sorghum mini core collection. *Theor. Appl. Genet.* 126, 2003–2015. doi: 10.1007/s00122-013-2113-x

Upadhyaya, H. D., Wang, L., Prakash, C. S., Liu, Y., Gao, L., Meng, R., et al. (2022). Genome-wide association mapping identifies an *SNF4* ortholog that impacts biomass and sugar yield in sorghum and sugarcane. *J. Exp. Bot.* 73, 3584–3596. doi: 10.1093/jxb/erac110

Valencia, R. C., and Rooney, W. L. (2009). *Genetic Control of Sorghum Grain Color* (San Andrés, La Libertad, El Salvador: Centro Nacional de Tecnologia Agropecuaria y Forestal/USAID/INTSORMIL). 10p.

Wang, Y. H., Upadhyaya, H. D., Burrell, A. M., Sahraeian, S. M., Klein, R. R., and Klein, P. E. (2013). Genetic structure and linkage disequilibrium in a diverse, representative collection of the C_4 model plant, *Sorghum bicolor*. *G3 (Bethesda)* 3, 783–793. doi: 10.1534/g3.112.004861

Wang, L., Upadhyaya, H. D., Zheng, J., Liu, Y., Singh, S. K., Gowda, C. L. L., et al. (2021). Genome-wide association mapping identifies novel panicle morphology loci and candidate genes in sorghum. *Front. Plant Sci.* 12, 743838. doi: 10.3389/fpls.2021.743838

Williams-Alanis, H., Rodriguez-Herrera, R., and Pecina-Quintero, V. (1999). Yield and agronomic traits relations with plant color of sorghum hybrids. *Cereal Res. Commun.* 27, 447–454. doi: 10.1007/BF03543562

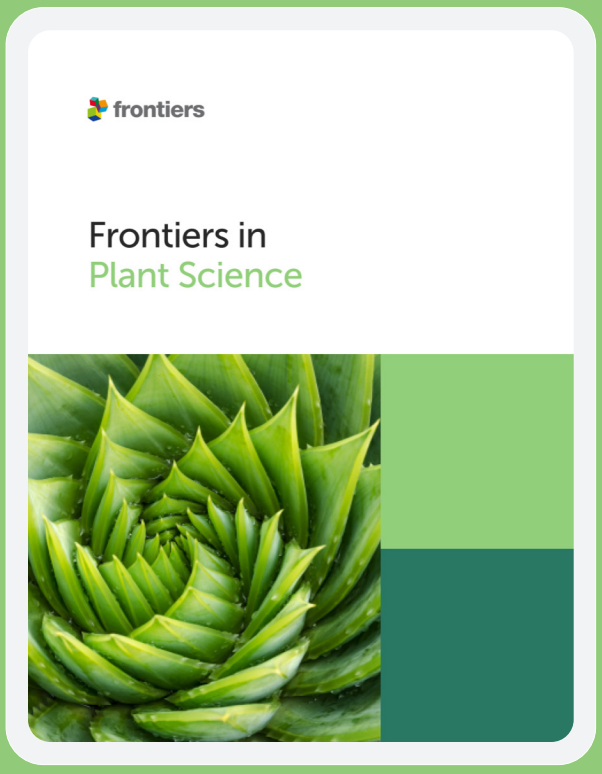
Frontiers in
Plant Science

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