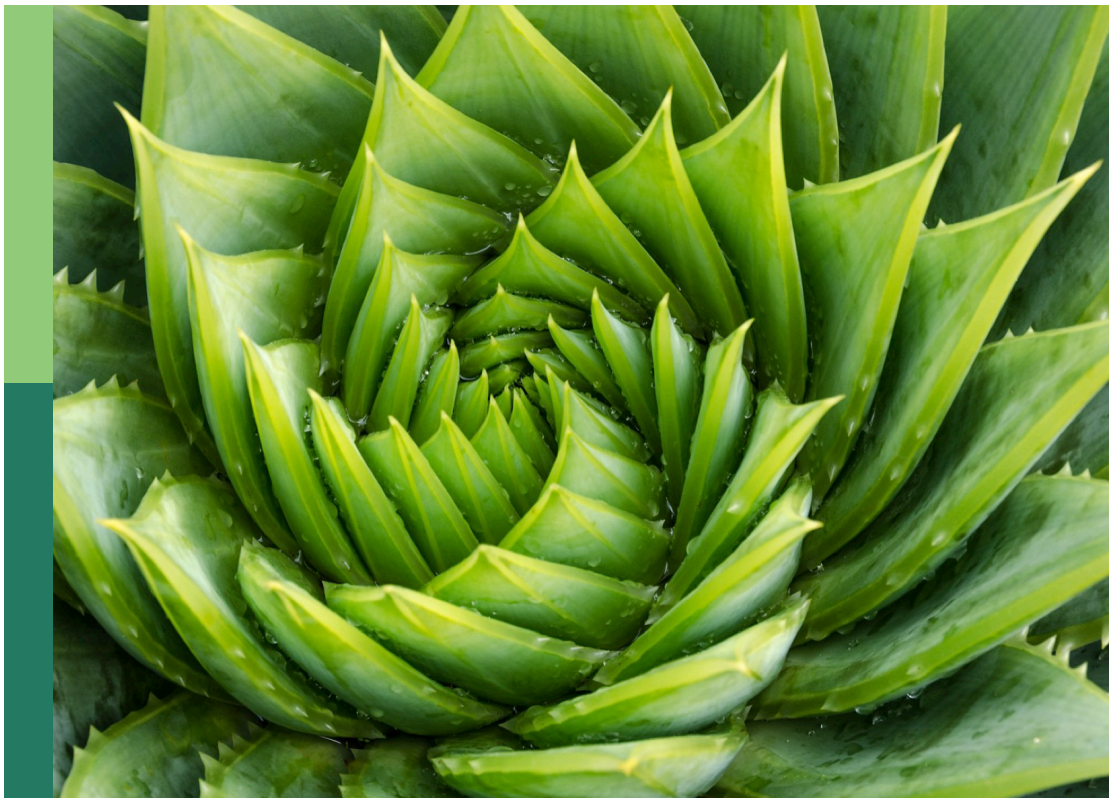
frontiers | Research Topics

Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces

Edited by
Andrés J. Cortés and Jinyoung Y. Barnaby

Published in
Frontiers in Plant Science
Frontiers in Genetics



ISSN 1664-8714
ISBN 978-2-83252-002-4
DOI 10.3389/978-2-83252-002-4

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

Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces

Topic editors

Andrés J. Cortés — Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

Jinyoung Y. Barnaby — United States Department of Agriculture (USDA), United States

Citation

Cortés, A. J., Barnaby, J. Y., eds. (2023). *Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83252-002-4

Table of contents

frontiers | Frontiers in Plant Science

TYPE Editorial
PUBLISHED 10 March 2023
DOI 10.3389/fpls.2023.1149469

OPEN ACCESS

EDITED AND REVIEWED BY
Andrea Zuccolo,
Kaust, Saudi Arabia

*CORRESPONDENCE
Andrés J. Cortés
acortes@agrosavia.co

†SECONDARY ADDRESS
Andrés J. Cortés,
Facultad de Ciencias Agrarias – Departamento de Ciencias Forestales, Universidad Nacional de Colombia –Sede Medellín, Medellín, Colombia

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

RECEIVED 22 January 2023
ACCEPTED 26 January 2023
PUBLISHED 10 March 2023

CITATION
Cortés AJ and Barnaby JY (2023) Editorial: Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces.
Front. Plant Sci. 14:1149469.
doi: 10.3389/fpls.2023.1149469

Editorial: Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces

Andrés J. Cortés[1*†] and Jinyoung Y. Barnaby[2]

[1]Corporación Colombiana de Investigación Agropecuaria – AGROSAVIA, C.I. La Selva, Rionegro, Colombia, [2]U.S. Department of Agriculture, U.S. National Arboretum, Floral and Nursery Plants Research Unit, Beltsville, MD, United States

KEYWORDS

germplasm, *ex situ* conservation, food security, pre-breeding, exotic variation, elite varieties, polygenic variation, adaptation

Editorial on the Research Topic
Harnessing genebanks: High-throughput phenotyping and genotyping of crop wild relatives and landraces

Introduction

Worldwide genebanks hold phenotypic and genetic novelty useful to increase yield, crop adaptability, and agrobiodiversity (Tanksley and McCouch, 1997) while buffering crop genetic erosion (Khoury et al., 2021). However, new strategies for genebank utilization must be empowered in order to meet increasing global food demand (McCouch, 2013; Bohra et al., 2021) with crop alternatives resilient to climate change, sustainable to the environment and the biodiversity, and profitable for communities (Scherer et al., 2020). Therefore, in order to contribute filling this gap on genebank mining, this Research Topic compiles recent developments able to speed up crop improvement processes by leveraging high-throughput phenotyping and genotyping of crop wild relatives (CWR) and landraces (Singh et al., 2022). As discussed in the next section, the amassed works innovate different steps of genebank characterization, utilization, and allelic deployment, including germplasm identification, conservation, pre-breeding screening for genepool diversity and associated markers, and introgression breeding.

Mining and unlocking the hidden value of CWR and landraces

Crop wild relatives are feasible sources for novel genetic and phenotypic variation (Bohra et al., 2021), as exemplified by Inagaki et al. The team demonstrated that a rice (*Oryza sativa* L.) near-isogenic line (NIL) that carries a genomic segment on chromosome seven from a Thai *O. rufipogon* CWR, narrowed by marker-assisted selection from a BC_4F_2 generation, is able to capture an optimum spectrum of plant shapes for sunlight receiving efficiency, which

in turn maximizes vegetative growth and weed suppression. The historical crop-wild transition and genetic interchange during domestication also reinforces the value of wild genepools, as shown by Yuan et al. The team recovered loci responsible for seed coat color pattern during domestication of wild soybean [*Glycine max* (L.) Merr.] through whole-genome re-sequencing of 276 F_{10} recombinant inbred lines (RILs) derived from a cross between cultivated and wild soybean accessions. Phenotypic variations from CWR may also be retained in the offspring between cultivated and wild genepools, as demonstrated by Li et al. (2021) Specifically, they captured 31 agronomic-relevant QTL and allelic variation (*e.g.*, for grain formation and length) from weedy rice (*Oryza sativa* f. *spontanea*) *via* whole-genome sequencing of a 199 F_2 population obtained by crossing a weedy rice with low heterozygosity and a cultivated rice variety. These results by Inagaki et al. (2021), Yuan et al., and Li et al. (2021) exemplify the utilization of wild genetic resources to increase the genetic base of cultivated crops for sustainable and resilient agricultural production.

Landraces, having been selected under subsistence agricultural environments, also hold a broad representation of useful natural variation (McCouch, 2004). For instance, Osterman et al. used LASSO models to link environmental variation with adaptive genetic diversity in red clover (*Trifolium pratense* L.), a perennial temperate forage legume, through genotyping 382 accessions from Scandinavia, including landraces and CWR, with 661 single nucleotide polymorphism (SNP) markers derived from seqSNP-targeted sequencing. The genomic landscape of genetic variation was also unveiled by Mulugeta et al. The authors used a diversity panel of durum wheat (*Triticum durum* Desf.) collected from the Ethiopian highlands, and identified 44 genomic regions associated with grain yield and related traits using last-generation (*i.e.*, FarmCPU) genome-wide association study (GWAS) models that inputted 10,045 SNP markers derived from the 25k Illumina wheat SNP array. In order to further explore isolated pockets of cryptic agrobiodiversity in Ethiopia, Enyew et al. and Gebeyehu et al. respectively examined the patterns of genetic diversity in sorghum [*Sorghum bicolor* (L.) Moench] and noug (*Guizotia abyssinica*), an outcrossing edible oilseed crop. The former team genotyped a total of 359 sorghum individuals, comprising 24 landrace accessions, with 3,001 gene-based single nucleotide polymorphism (SNP) markers, and found that sorghum accessions with bent peduncles exhibited more genetic variation than those with erect peduncles (Enyew et al.). The latter team performed RNA-Seq based transcriptome sequencing of 30 noug genotypes (Gebeyehu et al.), and predictably captured greater genetic similarity among self-compatible genotypes compared to self-incompatible accessions. Finally, Nuraga et al. introduced enset (*Ensete ventricosum*), a multipurpose crop grown in southern Ethiopia for human food, animal feed, and fiber. They demonstrated that enset landraces were not genetically differentiated based on their use-value (*i.e.*, medicinal *vs.* non-medicinal landraces) through genotyping 51 accessions with 15 simple sequence repeat (SSR) markers. Altogether, the works by Osterman et al. (2022), Enyew et al. (2022), Gebeyehu et al., and Nuraga et al. demonstrate the potential of landraces as sources of genetic innovation for agrobiodiversity.

The latter study by Nuraga et al. also illustrates how CWR and landraces are being used for exotic non-conventional purposes (Von Wettberg et al., 2020). Following Nuraga et al., medicinal applications were also addressed in the work by Cui el al., using Gardenia (*Gardenia jasminoides* Ellis), a Chinese perennial shrub from the Rubiaceae family with edible flowers and medicinal fruits. The authors conducted syntenic analyses with model species within the Rubiaceae family (*i.e.*, *Coffea arabica*, *C. canephora*, and *Ophiorrhiza pumila*), and discovered 18 and 31 QTL associated with growth- and leaf-related traits, respectively, by genotyping 200 F_1 hybrids with 4,249 genotyping by sequencing (GBS)-derived SNP markers. The couple of studies by Nuraga et al. and Cui et al. provide fundamental knowledge that enables further exploration of alternative uses of CWR and landraces (Wu et al., 2022).

Meanwhile, hybridization has played a key role in bidirectional adaptive introgression (Abbott et al., 2013) as well as hybrid breeding within crop-wild complexes (Cortés et al., 2022a). For example, Volk et al. (2022) explored the signatures of admixture from the cultivated apple [*Malus domestica* (Suckow) Borkh] into the progenitor species *M. sieversii* (Ledeb.) M. Roem., *M. orientalis* Uglitzk., and *M. sylvestris* (L.) Mill. (Cornille et al., 2014) by genotyping 463 accessions using the 20K apple SNP array. On the other hand, Conejo-Rodriguez et al. leveraged hybrid phenotyping for pre-breeding in beans (*Phaseolus* spp.). They proposed an innovative pipeline to determine parental phenomic proportions in hybrids, and validated it using interspecific crosses derived from common bean (*P. vulgaris* L.), tepary bean (*P. acutifolius* A. Gray), and its wild relative *P. parvifolius* Freytag. The latter two served as donors of heat and drought tolerance (tepary, Buitrago-Bitar et al., 2021), and resistance to common bacterial blight (CBB), respectively (Conejo-Rodriguez et al.). Both studies by Volk et al. (2022) and Conejo-Rodriguez et al. (2022) highlight introgression breeding as a still valid alternative to bridge historical barriers between crops and their wild genepools (Labroo et al., 2021).

An additional study harnessing modern phenomic 'big data' tools in diverse genepools was reported by Restrepo-Montoya et al. They utilized qualitative descriptors to curate and catalogue cotton (*Gossypium* spp.) germplasm encompassing an impressive dataset of 1,616 observations between 2011 and 2019 on 7,941 unique accessions and 50 species. The last two pre-breeding phenotyping efforts carried out by Conejo Rodriguez et al. and Restrepo-Montoya et al. enable designing and targeting downstream goals in breeding by defining trait families, categorizing germplasm diversity, and pinpointing exotic outlier accessions.

Although great advancement has been made in the field of phenotyping, matching the exponential achievement from the genomic arena still requires more innovative analytical tools. Candidate platforms should be capable to merge and interpret complex multi-dimensional datasets embracing diversity from CWR and landraces. To this end, the review by Tirnaz et al. (2022) envisioned novel avenues to utilize genebank resources, for instance *via de novo* domestication, genome editing, and speed breeding. This review further advocated for computational (*e.g.*, machine learning) approaches suitable for speeding up the incorporation of exotic genepools into breeding programs to improve crop production, adaptability and sustainability (Tirnaz et al.).

A roadmap to harness genebanks

Crop wild relatives and landraces stored at genebanks harbor unique variation that may benefit food security, sustainability (Tanksley and McCouch, 1997) and resilient climate change adaptation (Renzi et al., 2022). Yet, their factual utilization has been hampered by poor characterizations, genetic incompatibilities, and polygenic variance (Coyne et al., 2020). Therefore, an improved roadmap to address these bottlenecks with modern technologies is a prerequisite for their effective utilization (Figure 1). A first pivotal research avenue worth considering is mining phenotypic and genetic variation hidden within CWR and landraces (Singh et al., 2022) through extended sampling targeting isolated pockets of cryptic diversity (Ramirez-Villegas et al., 2020), robust ecological data curation (Waldvogel et al., 2020b) [*e.g.*, targeting specific abiotic stresses such as drought (Cortés and Blair, 2018) and heat tolerance (López-Hernández and Cortés)], dense linkage disequilibrium (LD) guided genomic characterizations (Blair et al., 2018), and geographic-wide agronomical (Osterman et al. 2022) and physiological (Conejo-Rodriguez et al.) trials across diverse germplasm and environments [recently referred to as enviromics (Costa-Neto and Fritsche-Neto, 2021; Crossa et al., 2021; Resende et al., 2021)].

Second, bridging crop–wild incompatibilities and unlocking novel diversity may rely on hybrid breeding *via* bridge genotypes (Conejo-Rodriguez et al. 2022), and introgression breeding (Migicovsky and Myles, 2017) from wild genepools into cultivars [*e.g.*, Inagaki et al. (2021), Yuan et al. (2022), and Li et al. (2021)], and viceversa (Volk et al.). To overcome genotype incompatibility and embryo abortion issues, typically found in simpler backcrosses, pre-breeding schemes could be advanced into multiple alternating backcrosses between exotic donors and elite genepools (Burbano-Erazo et al., 2021), as in classical congruity backcrossing (Muñoz et al., 2003). Other promising alternatives to achieve a more optimal retention of the exotic phenotype include: embryo rescue, genomic-assisted backcrossing (Migicovsky and Myles, 2017), speed breeding (Watson et al., 2018; Alves et al., 2020; Bohra et al., 2021), tissue-specific gene editing [*e.g.*, for exotic alleles (Martignago et al., 2019), *de novo* domestication (Lemmon et al., 2018), precocious flowering time, and compatibility factors (Fritsche et al., 2018)], and grafting with wild rootstocks [*e.g.*, for biotic and abiotic stress tolerance (Warschefsky et al., 2016), adaptation to soil toxicity (Fernández-Paz et al., 2021), growth traits (Cañas-Gutiérrez et al., 2022), and yield (Reyes-Herrera et al., 2020)]. Merging these strategies could facilitate even quicker transitions from the wild genepools compared to classical inter-specific and crop-wild controlled pollination.

As a third step, major challenges to utilize CWR and landraces for improvement of complex polygenic adaptive traits (Renzi et al., 2022) can be tackled by linking last-generation experimental setups [*e.g.*, diversifying selection (McCouch, 2004), introgression breeding (Muñoz et al., 2003; Burgarella et al., 2019), speed breeding (Watson et al., 2018; Alves et al., 2020; Kumar et al., 2020; Varshney et al., 2021b), *de novo* domestication (Fernie and Yan, 2019), and genome editing (Lemmon et al., 2018)] with modern analytical developments [*e.g.*, last-generation genetic clustering (Enyew et al., 2022; Gebeyehu et al., 2022; Nuraga et al., 2022) and mapping Mulugeta et al. 2023, predictive breeding (Ahmar et al., 2021), genomic-informed selection (Desta and Ortiz, 2014; Crossa et al., 2017), and machine learning (Varshney, 2021; Tirnaz et al., 2022)]. Such trans-disciplinary approaches (Tirnaz et al., 2022) will enable better reconstructions of the genomic landscapes (Ellegren and Wolf, 2017; Barnaby et al., 2020; Tong and Nikoloski, 2021) of trait variation (Barnaby et al., 2022; Li et al.; Yuan et al.), and the selection signatures of past (Cortés et al., 2020; Waldvogel et al., 2020) and modern adaptation (Cortés et al., 2022b; Osterman et al., 2022). Retrieving complex polygenic interactions across the genomic architectures of quantitative (Boyle et al., 2017) and adaptive traits (Barrett and Hoekstra, 2011; Barghi et al., 2020) is a prerequisite to pinpoint exotic alleles enclosed within CWR and landraces (Cortés et al., 2012), and speed up the utilization of this natural variation as part of conservation and pre-breeding programs (Migicovsky and Myles, 2017).

Fourth, novel approaches need to de defined in order to leverage CWR and landraces (Varshney et al., 2021b; Tirnaz et al.). After all, boosting crop adaptability, sustainability, and yield in the face of changing climate, biodiversity loss, and increased pollution requires multilateral trans-disciplinary efforts. Various actors, including producers, scientist and transfer officers, decision makers, funding bodies and marketers should converge to plan and implement innovative strategies capable to meet future global food demands, while facing enlarged threats from abiotic (Blair et al., 2016; Cortés and López-Hernández, 2021) and biotic (Guevara-Escudero et al., 2021) pressures. Some of this pioneering strategies may include (1) prospection of genotypes candidate for exotic alleles (Migicovsky and Myles, 2017), ancient cultivars (Atchison et al., 2016; Thapa et al., 2021), and novel crops (Von Wettberg et al., 2020; Cui et al., 2022; Nuraga et al.); (2) multidimensional 'big data' compilation and management from the scientific, climatic and marketing arenas (Tirnaz et al., 2022) to identify conservation priorities (Castaneda-Alvarez et al., 2016), novel markets and value chains (Smale and Jamora, 2020), and (3) innovation for the production, transformation, delivery, commercialization and utilization of CWR- and landrace-derived crop varieties (McCouch, 2013) and ecosystem services (Tyack et al., 2020).

Perspectives

Further studies and implementation practices are urgently required to better harness high-throughput phenotyping and genotyping of CWR and landraces at genebanks. Ultimately, CWR and landraces can be utilized as resources to effectively explore phenotypic and genetic variants associated with innovative phenotypes, and to transfer those traits into customized cultivars (Varshney et al., 2021a). Alternatively, CWR and landraces may themselves be adopted as novel food sources (Von Wettberg et al., 2020). For these strategies to succeed, holistic multilateral and trans-disciplinary agendas should also prioritize an open-access networks approach (Spindel and McCouch, 2016) for mobilizing crop biodiversity (McCouch et al., 2016). On balance, embracing food security with resilient and sustainable crop systems requires equalizing multidimensional access of communities, minorities and

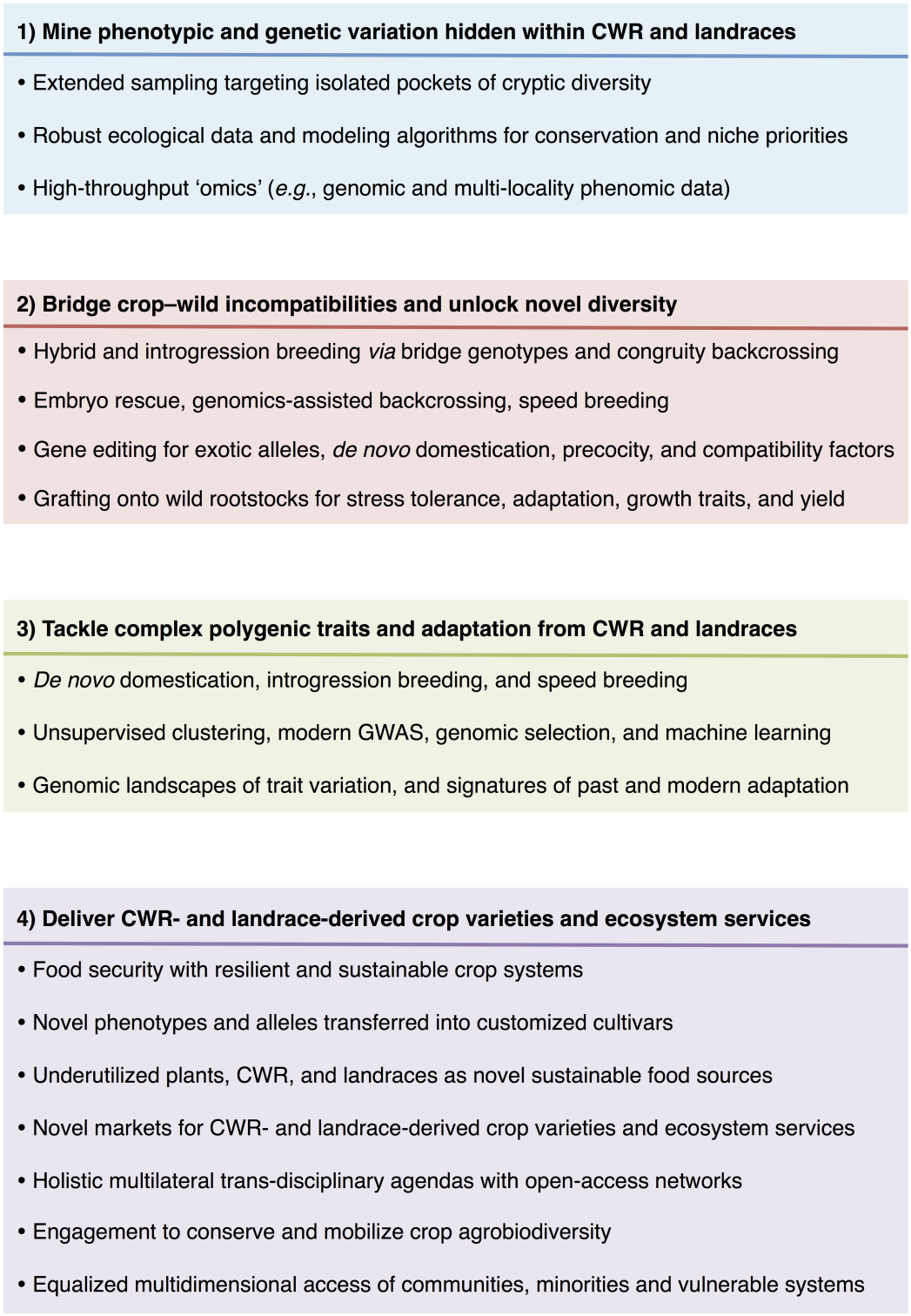


FIGURE 1
Schematic roadmap to harness genebanks. Each box marks pivotal research avenues (with the corresponding recommendations within) worth considering for mining, unlocking and utilizing phenotypic and genetic novelty hidden within crop wild relatives (CWR) and landraces. Ultimately, achieving food security with resilient and sustainable agrifood systems capable to cope with climate change, biodiversity loss, and increased pollution would require nature-based solutions for climate-smart alimentary schemes inspired in underutilized plant species, CWR, and landraces.

vulnerable systems, including farmers from developing countries, women in agriculture, early-career researchers, global South-South partners, and orphan crops. Only then, underutilized plants, CWR and landraces would play an active role as nature-based solutions for agrifood systems.

Author contributions

AC conceived the Research Topic with insights from JB. Both authors made substantial contributions in preparing, editing and reviewing the contents of the Research Topic on "*Harnessing Genebanks: High-Throughput Phenotyping and Genotyping of CWR and Landraces*". AC wrote a first version of the editorial, later edited by JB. Both approved it for publication.

Funding

Vetenskapsrådet (grants 2016-00418 and 2022-04411), Kungliga Vetenskapsakademien (grant BS20170036), and the British Council (grant 527023146) are deeply acknowledged for funding AJC as PI during the conception and execution of this Research Topic. The Graduate Research School in Genomic Ecology (GENECO) from Lund University is also recognized for the mobility funding that assisted the synergistic meeting between AC and M.W. Blair in the spring of 2015 at Nashville, TN, United States. Fulbright U.S. Specialist Program is thanked for supporting M.W. Blair's visit to AC in Rionegro (Antioquia, Colombia) in the summer of 2019. Both encounters ultimately laid the conceptual foundations of this Research Topic.

Acknowledgments

The authors express their profound acknowledgment to all contributors, including authors, reviewers and assisting editors, who made possible this successful Research Topic. Discussions with M.W. Blair, I. Cerón-Souza, C.H. Galeano, F. López-Hernández, C.I. Medina, A.A. Navas-Arboleda, D. Peláez, P.H. Reyes-Herrera, A.P. Tofiño-Rivera, M. Urban, and R. Yockteng where insightful for some of the perspectives on genebank utilization discussed here. Recognition is also given to M.J. Torres for assistance during the preparation of this Research Topic. Finally, we are as well in debt with the editorial board of Frontiers in Plant Science and its staff for encouraging, assisting and enabling this Research Topic on "*Harnessing Genebanks: High-Throughput Phenotyping and Genotyping of Crop Wild Relatives and Landraces*".

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

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., et al. (2013). Hybridization and speciation. *J. Evolutionary Biol.* 26, 229–246. doi: 10.1111/j.1420-9101.2012.02599.x

Ahmar, S., Ballesta, P., Ali, M., and Mora-Poblete, F. (2021). Achievements and challenges of genomics-assisted breeding in forest trees: From marker-assisted selection to genome editing. *Int. J. Mol. Sci.* 22, 10583. doi: 10.3390/ijms221910583

Alves, F. C., Balmant, K. M., Resende, M. F. R., Kirst, M., and Los Campos, G. (2020). Accelerating forest tree breeding by integrating genomic selection and greenhouse phenotyping. *Plant Genome* 13, e20048. doi: 10.1002/tpg2.20048

Atchison, G. W., Nevado, B., Eastwood, R. J., Contreras-Ortiz, N., Reynel, C., Madrinan, S., et al. (2016). Lost crops of the incas: Origins of domestication of the Andean pulse crop tarwi, *Lupinus mutabilis*. *Am. J. Bot.* 103, 1592–1606. doi: 10.3732/ajb.1600171

Barnaby, J. Y., Huggins, T. D., Lee, H., McClung, A. M., Pinson, S. R. M., Oh, M., et al. (2020). Vis/NIR hyperspectral imaging distinguishes sub-population, production environment, and physicochemical grain properties in rice. *Sci. Rep.* 10, 9284. doi: 10.1038/s41598-020-65999-7

Barnaby, J. Y., McClung, A. M., Edwards, J. D., Pinson, S. R. M., et al. (2022). Identification of quantitative trait loci for tillering, root, and shoot biomass at the maximum tillering stage in rice. *Sci. Rep.* 12, 13304. doi: 10.1038/s41598-022-17109-y

Barghi, N., Hermisson, J., and SchlöTterer, C. (2020). Polygenic adaptation: A unifying framework to understand positive selection. *Nat. Rev. Genet.* 21, 769–781. doi: 10.1038/s41576-020-0250-z

Barrett, R. D. H., and Hoekstra, H. E. (2011). Molecular spandrels: Tests of adaptation at the genetic level. *Nat. Rev. Genet.* 12, 767–780. doi: 10.1038/nrg3015

Blair, M. W., Cortes, A. J., Farmer, A. D., Huang, W., Ambachew, D., Penmetsa, R. V., et al. (2018). Uneven recombination rate and linkage disequilibrium across a reference snp map for common bean (*Phaseolus vulgaris* L.). *PloS One* 13, e0189597. doi: 10.1371/journal.pone.0189597

Blair, M. W., Cortés, A. J., and This, D. (2016). Identification of an *Erecta* gene and its drought adaptation associations with wild and cultivated common bean. *Plant Sci.* 242, 250–259. doi: 10.1016/j.plantsci.2015.08.004

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., Mccouch, S. R., et al. (2021). Reap the crop wild relatives for breeding future crops. *Trends Biotechnol* 40(4), 412-431. doi: 10.1016/j.tig.2021.08.002

Boyle, E. A., Li, Y. I., and Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169, 1177–1186. doi: 10.1016/j.cell.2017.05.038

Buitrago-Bitar, M. A., Cortés, A. J., López-Hernández, F., Londoño-Caicedo, J. M., Muñoz-Florez, J. E., Muñoz, L. C., et al. (2021). Allelic diversity at abiotic stress responsive genes in relationship to ecological drought indices for cultivated tepary bean, *Phaseolus acutifolius* A. Gray, and its wild relatives. *Genes* 12, 556. doi: 10.3390/genes12040556

Burbano-Erazo, E., León-Pacheco, R., Cordero-Cordero, C., López-Hernández, F., Cortés, A., and Tofiño-Rivera, A. (2021). Multi-environment yield components in advanced common bean (*Phaseolus vulgaris* L.) × tepary bean (*P. acutifolius* A. Gray) interspecific lines for heat and drought tolerance. *Agronomy* 11, 1978. doi: 10.3390/agronomy11101978

Burgarella, C., Barnaud, A., Kane, N. A., Jankowski, F., Scarcelli, N., Billot, C., et al. (2019). Adaptive introgression: An untapped evolutionary mechanism for crop adaptation. *Front. Plant Sci.* 10, 4. doi: 10.3389/fpls.2019.00004

Cañas-Gutiérrez, G. P., Sepulveda-Ortega, S., López-Hernández, F., Navas-Arboleda, A. A., and Cortés, A. J. (2022). Inheritance of yield components and morphological traits in avocado cv. Hass from "Criollo" "Elite trees" *Via* half-Sib seedling rootstocks. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.843099

Castaneda-Alvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., et al. (2016). Global conservation priorities for crop wild relatives. *Nat. Plants* 2, 16022. doi: 10.1038/nplants.2016.22

Cornille, A., Giraud, T., Smulders, M. J. M., Roldán-Ruiz, I., and Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends Genet.* 30, 57–65. doi: 10.1016/j.tig.2013.10.002

Cortés, A. J., and López-Hernańndez, F. (2021). Harnessing crop wild diversity for climate change adaptation. *Genes* 12, 783. doi: 10.3390/genes12050783

Cortés, A. J., and Blair, M. W. (2018). Genotyping by sequencing and genome – environment associations in wild common bean predict widespread divergent adaptation to drought. *Front. Plant Sci.* 9, 128. doi: 10.3389/fpls.2018.00128

Cortés, A. J., Cornille, A., and Yockteng, R. (2022a). Evolutionary genetics of crop-wild complexes. *Genes* 13, 1. doi: 10.3390/genes13010001

Cortés, A. J., López-Hernández, F., and Blair, M. W. (2022b). Genome–environment associations, an innovative tool for studying heritable evolutionary adaptation in orphan crops and wild relatives. *Front. Genet.* 13, 910386. doi: 10.3389/fgene.2022.910386

Cortés, A. J., López-Hernández, F., and Osorio-Rodriguez, D. (2020). Predicting thermal adaptation by looking into populations' genomic past. *Front. Genet.* 11, 564515. doi: 10.3389/fgene.2020.564515

Cortés, A. J., This, D., Chavarro, C., Madriñán, S., and Blair, M. W. (2012). Nucleotide diversity patterns at the drought-related *Dreb2* encoding genes in wild and cultivated common bean (*Phaseolus vulgaris* L.). *Theor. Appl. Genet.* 125, 1069–1085. doi: 10.1007/s00122-012-1896-5

Costa-Neto, G., and Fritsche-Neto, R. (2021). Enviromics: Bridging different sources of data, building one framework. *Crop Breed. Appl. Biotechnol.* 21, e393521S393512. doi: 10.1590/1984-70332021v21sa25

Coyne, C. J., Kumar, S., Von Wettberg, E. J. B., Marques, E., Berger, J. D., Redden, R. J., et al. (2020). Potential and limits of exploitation of crop wild relatives for pea, lentil, and chickpea improvement. *Legume Sci.* 2 (2), e36. doi: 10.1002/leg3.36

Crossa, J., Fritsche-Neto, R., Montesinos-Lopez, O. A., Costa-Neto, G., Dreisigacker, S., Montesinos-Lopez, A., et al. (2021). The modern plant breeding triangle: Optimizing the use of genomics, phenomics, and enviromics data. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.651480

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

Desta, Z. A., and Ortiz, R. (2014). Genomic selection: Genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006

Ellegren, H., and Wolf, J. B. W. (2017). Parallelism in genomic landscapes of differentiation, conserved genomic features and the role of linked selection. *J. Evol. Biol.* 30, 1516–1518. doi: 10.1111/jeb.13113

Fernández-Paz, J., Cortés, A. J., Hernández-Varela, C. A., Mejía-De-Tafur, M. S., Rodriguez-Medina, C., and Baligar, V. C. (2021). Rootstock-mediated genetic variance in cadmium uptake by juvenile cacao (*Theobroma cacao* L.) genotypes, and its effect on growth and physiology. *Front. Plant Sci.* 12. 10.3389/fpls.2021.777842

Fernie, A. R., and Yan, J. (2019). *De novo* domestication: An alternative route toward new crops for the future. *Mol. Plant* 12, 615–631. doi: 10.1016/j.molp.2019.03.016

Fritsche, S., Klocko, A. L., Boron, A., Brunner, A. M., and Thorlby, G. (2018). Strategies for engineering reproductive sterility in plantation forests. *Front. Plant Sci.* 9, 1671. doi: 10.3389/fpls.2018.01671

Guevara-Escudero, M., Osorio, A. N., and Cortés, A. J. (2021). Integrative pre-breeding for biotic resistance in forest trees. *Plants* 10, 2022. doi: 10.3390/plants10102022

Inagaki, N., Asami, H., Hirabayashi, H., Uchino, A., Imaizumi, T., and Ishimaru, K. (2021). A rice ancestral genetic resource conferring ideal plant shapes for vegetative growth and weed suppression. *Front. Plant Sci.* 12, 748531. doi: 10.3389/fpls.2021.748531

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., Haan, S., Engels, J. M. M., et al. (2021). Crop genetic erosion: Understanding and responding to loss of crop diversity. *New Phytol.* 233, 84–118. doi: 10.1111/nph.17733

Kumar, S., Hilario, E., Deng, C. H., and Molloy, C. (2020). Turbocharging introgression breeding of perennial fruit crops: A case study on apple. *Horticulture Res.* 7, 47. doi: 10.1038/s41438-020-0270-z

Labroo, M. R., Studer, A. J., and Rutkoski, J. E. (2021). Heterosis and hybrid crop breeding: A multidisciplinary review. *Front. Genet.* 12. doi: 10.3389/fgene.2021.643761

Lemmon, Z. H., Reem, N. T., Dalrymple, J., Soyk, S., Swartwood, K. E., Rodriguez-Leal, D., et al. (2018). Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* 4, 766–770. doi: 10.1038/s41477-018-0259-x

López-Hernández, F., and Cortés, A. J. (2019). Last-generation genome–environment associations reveal the genetic basis of heat tolerance in common bean (*Phaseolus vulgaris* L.). *Front. Genet.* 10, 22. doi: 10.3389/fgene.2019.00954

Martignago, D., Rico-Medina, A., Blasco-Escamez, D., Fontanet-Manzaneque, J. B., and Cano-Delgado, A. I. (2019). Drought resistance by engineering plant tissue-specific responses. *Front. Plant Sci.* 10, 1676. doi: 10.3389/fpls.2019.01676

McCouch, S. (2004). Diversifying selection in plant breeding. *PloS Biol.* 2, 1507–1512. doi: 10.1371/journal.pbio.0020347

McCouch, S. (2013). Feeding the future. *Nature* 499, 23–24. doi: 10.1038/499023a

McCouch, S. R., Wright, M. H., Tung, C. W., Maron, L. G., Mcnally, K. L., Fitzgerald, M., et al. (2016). Open access resources for genome-wide association mapping in rice. *Nat. Commun.* 7, 10532. doi: 10.1038/ncomms10532

Migicovsky, Z., and Myles, S. (2017). Exploiting wild relatives for genomics-assisted breeding of perennial crops. *Front. Plant Sci.* 8, 460. doi: 10.3389/fpls.2017.00460

Muñoz, L. C., Blair, M. W., Duque, M. C., Tohme, J., and Roca, W. (2003). Introgression in common bean X tepary bean interspecific congruity-backcross lines as measured by aflp markers. *Crop Sci.* 44, 637–645. doi: 10.2135/cropsci2004.6370

Ramirez-Villegas, J., Khoury, C. K., Achicanoy, H. A., Mendez, A. C., Diaz, M. V., Sosa, C. C., et al. (2020). A gap analysis modelling framework to prioritize collecting for ex situ conservation of crop landraces. *Divers. Distrib* 26, 730–742. doi: 10.1111/ddi.13046

Renzi, J. P., Coyne, C. J., Berger, J., Von Wettberg, E., Nelson, M., Ureta, S., et al. (2022). How could the use of crop wild relatives in breeding increase the adaptation of crops to marginal environments? *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.886162

Resende, R. T., Piepho, H. P., Rosa, G. J. M., Silva-Junior, O. B., Silva, F. F. E., Resende, M. D. V. D., et al. (2021). *Enviromics* in breeding: Applications and perspectives on envirotypic-assisted selection. *Theor. Appl. Genet.* 134, 95–112. doi: 10.1007/s00122-020-03684-z

Reyes-Herrera, P. H., Muñoz-Baena, L., Velásquez-Zapata, V., Patiño, L., Delgado-Paz, O. A., Díaz-Diez, C. A., et al. (2020). Inheritance of rootstock effects in avocado (*Persea americana* mill.) cv. hass. *Front. Plant Sci.* 11, 555071. doi: 10.3389/fpls.2020.555071

Scherer, L., Svenning, J. C., Huang, J., Seymour, C. L., Sandel, B., Mueller, N., et al. (2020). Global priorities of environmental issues to combat food insecurity and biodiversity loss. *Sci. Total Environ.* 730, 139096. doi: 10.1016/j.scitotenv.2020.139096

Singh, G., Gudi, S., Amandeep, Upadhyay, P., Shekhawat, P. K., Nayak, G., Goyal, L., et al. (2022). Unlocking the hidden variation from wild repository for accelerating genetic gain in legumes. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1035878

Smale, M., and Jamora, N. (2020). Valuing genebanks. *Food Secur.* 12, 905–918. doi: 10.1007/s12571-020-01034-x

Spindel, J. E., and McCouch, S. R. (2016). When more is better: How data sharing would accelerate genomic selection of crop plants. *New Phytol.* 212, 814–826. doi: 10.1111/nph.14174

Tanksley, S. D., and McCouch, S. R. (1997). Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* 227, 1063–1066. doi: 10.1126/science.277.5329.1063

Thapa, R., Edwards, M., and Blair, M. W. (2021). Relationship of cultivated grain amaranth species and wild relative accessions. *Genes* 12, 1849. doi: 10.3390/genes12121849

Tong, H., and Nikoloski, Z. (2021). Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *J. Plant Physiol.* 257, 153354. doi: 10.1016/j.jplph.2020.153354

Tyack, N., Dempewolf, H., and Khoury, C. K. (2020). The potential of payment for ecosystem services for crop wild relative conservation. *Plants* 9, 1305. doi: 10.3390/plants9101305

Varshney, R. K. (2021). The plant genome special issue: Advances in genomic selection and application of machine learning in genomic prediction for crop improvement. *Plant Genome* 14(3), e20178. doi: 10.1002/tpg2.20178

Varshney, R. K., Barmukh, R., Roorkiwal, M., Qi, Y., Kholova, J., Tuberosa, R., et al. (2021a). Breeding custom-designed crops for improved drought adaptation. *Advanced Genet.* 2, e202100017. doi: 10.1002/ggn2.202100017

Varshney, R. K., Bohra, A., Roorkiwal, M., Barmukh, R., Cowling, W. A., Chitikineni, A., et al. (2021b). Fast-forward breeding for a food-secure world. *Trends Genet.* 37, 1124–1136. doi: 10.1016/j.tig.2021.08.002

Von Wettberg, E., Davis, T. M., and Smýkal, P. (2020). Wild plants as source of new crops. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.591554

Waldvogel, A. M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., et al. (2020a). Evolutionary genomics can improve prediction of species' responses to climate change. *Evol. Lett.* 4, 4–18. doi: 10.1002/evl3.154

Waldvogel, A. M., Schreiber, D., Pfenninger, M., and Feldmeyer, B. (2020b). Climate change genomics calls for standardised data reporting. *Front. Ecol. Evol.* 8, 242. doi: 10.3389/fevo.2020.00242

Warschefsky, E. J., Klein, L. L., Frank, M. H., Chitwood, D. H., Londo, J. P., Von Wettberg, E. J. B., et al. (2016). Rootstocks: Diversity, domestication, and impacts on shoot phenotypes. *Trends Plant Sci.* 21, 418–437. doi: 10.1016/j.tplants.2015.11.008

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M. D., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* 4, 23–29. doi: 10.1038/s41477-017-0083-8

Wu, X., Cortés, A. J., and Blair, M. W. (2022). Genetic differentiation of grain, fodder and pod vegetable type cowpeas (*Vigna unguiculata* L.) identified through single nucleotide polymorphisms from genotyping-by-Sequencing. *Mol. Horticulture* 2, 8. doi: 10.1186/s43897-022-00028-x

frontiers in Plant Science

ORIGINAL RESEARCH
published: 19 November 2021
doi: 10.3389/fpls.2021.775051

High-Quality Genomes and High-Density Genetic Map Facilitate the Identification of Genes From a Weedy Rice

Fei Li[1†], Zhenyun Han[1†], Weihua Qiao[1], Junrui Wang[1,2], Yue Song[1], Yongxia Cui[1,3], Jiaqi Li[1,4], Jinyue Ge[1], Danjing Lou[1], Weiya Fan[1], Danting Li[5], Baoxuan Nong[5], Zongqiong Zhang[5], Yunlian Cheng[1], Lifang Zhang[1], Xiaoming Zheng[1] and Qingwen Yang[1]**

[1] National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] Guangxi Key Laboratory for Polysaccharide Materials and Modifications, School of Marine Sciences and Biotechnology, Guangxi University for Nationalities, Nanning, China, [3] School of Clinical Medicine, Southwest Medical University, Luzhou, China, [4] Little Berry Research Room, Liaoning Institute of Fruit Science, Yingkou, China, [5] Guangxi Key Laboratory of Rice Genetics and Breeding, Rice Research Institute, Guangxi Academy of Agricultural Sciences, Nanning, China

OPEN ACCESS

Edited by:
Andrés J. Cortés, Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

Reviewed by:
Maria Fernanda Alvarez, Rice Program International Centre for Tropical Agriculture (CIAT), Colombia
Joong Hyoun Chin, Sejong University, South Korea

***Correspondence:**
Xiaoming Zheng
zhengxiaoming@caas.cn
Qingwen Yang
yangqingwen@caas.cn

† These authors have contributed equally to this work

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

Received: *13 September 2021*
Accepted: *27 October 2021*
Published: *19 November 2021*

Citation:
Li F, Han Z, Qiao W, Wang J, Song Y, Cui Y, Li J, Ge J, Lou D, Fan W, Li D, Nong B, Zhang Z, Cheng Y, Zhang L, Zheng X and Yang Q (2021) High-Quality Genomes and High-Density Genetic Map Facilitate the Identification of Genes From a Weedy Rice. Front. Plant Sci. 12:775051. doi: 10.3389/fpls.2021.775051

Genes have been lost or weakened from cultivated rice during rice domestication and breeding. Weedy rice (*Oryza sativa* f. *spontanea*) is usually recognized as the progeny between cultivated rice and wild rice and is also known to harbor an gene pool for rice breeding. Therefore, identifying genes from weedy rice germplasms is an important way to break the bottleneck of rice breeding. To discover genes from weedy rice germplasms, we constructed a genetic map based on w-hole-genome sequencing of a F_2 population derived from the cross between LM8 and a cultivated rice variety. We further identified 31 QTLs associated with 12 important agronomic traits and revealed that *ORUFILM03g000095* gene may play an important role in grain length regulation and participate in grain formation. To clarify the genomic characteristics from weedy rice germplasms of LM8, we generated a high-quality genome assembly using single-molecule sequencing, Bionano optical mapping, and Hi-C technologies. The genome harbored a total size of 375.8 Mb, a scaffold N50 of 24.1 Mb, and originated approximately 0.32 million years ago (Mya) and was more closely related to *Oryza sativa* ssp. *japonica*. and contained 672 unique genes. It is related to the formation of grain shape, heading date and tillering. This study generated a high-quality reference genome of weedy rice and high-density genetic map that would benefit the analysis of genome evolution for related species and suggested an effective way to identify genes related to important agronomic traits for further rice breeding.

Keywords: weedy rice, genetic map, QTL mapping, reference genome, comparative genomics

INTRODUCTION

Cultivated rice is one of the most important staple crops worldwide. The breeding of rice varieties with improved yield, quality, resistance to diseases and pests, and tolerance to abiotic stresses is significant to meet the increasing food demand in China and the world (Khush, 2001; Yang and Hwa, 2008; Xu et al., 2021). However, many genes have been lost from cultivated rice due

to the long-term domestication and artificial selection, which hinders the breeding of advanced rice varieties. To the contrary, wild rice growing in natural environments is resistant or tolerant to different biotic and abiotic stresses and therefore retains a natural gene pool containing a large number of genes that have been lost or weakened from cultivated rice (Sun et al., 2002). Weedy rice has many characteristic traits similar to those of wild rice, many studies indicated that weedy rice was originated from wild rice and serves as a transition type between wild rice and cultivated rice (Baker, 1974; Wet and Harlan, 1975; Cho et al., 1995). Previous studies showed that weedy rice harbors the AA genome and no reproductive isolation was observed between weedy rice and cultivated rice (Nadir et al., 2018; Sun et al., 2019). Generally, the genes of weedy rice can be transferred to cultivated rice through breeding techniques such as hybridization and backcrossing (Lu et al., 2000; Stein et al., 2018). Weedy rice has been usually used as the genetic materials for rice genetics and breeding or to identify genes related to stress tolerance, disease and pest resistance, high yield, and high grain quality for improving modern rice varieties (Ishikawa et al., 2005; Shivrain et al., 2010; Dai et al., 2013).

In the past decades, rice functional genomics research, which focuses on technology platform construction and molecular cloning and functional analysis of genes related to important agronomic traits, has resulted in numerous achievements in gene discovery (Han et al., 2007; Xu et al., 2021). Due to its small genome and relatively simple structure, *Oryza sativa* (9311 and Nipponbare) became the first sequenced rice species in 2002 (Goff et al., 2002; Yu et al., 2002). These rice reference genomes have enabled massive rice functional genomics research, accelerated rice genetic improvement, and laid a foundation for studying genomes of other crops such as *Zea mays* (Schnable et al., 2009) and *Triticum aestivum* (International Wheat Genome Sequencing Consortium [IWGSC], 2014). With the development of sequencing technology, the time required for sequencing has largely decreased while the sequencing quality has greatly improved, therefore resulting in more high-quality reference genomes of cultivated rice varieties such as MH63, ZH97, and R498 (Zhang et al., 2016, 2018; Du et al., 2017). The focus of rice research has also been gradually turned to elucidate biological characteristics and evolution processes and to analyze gene functions and related biological issues at the genomic level, as well as to identify genes related to important agronomic traits such as high yield, high quality, and stress resistance (Huang et al., 2010, 2011, 2015; Xun et al., 2012; Wei et al., 2014; Yano et al., 2016).

At present, numerous genes related to important agronomic traits (e.g., grain size) have been located and cloned, such as GS3 (Fan et al., 2006; Mao et al., 2018), GL3.1 (Qi et al., 2012), DEP1 (Huang et al., 2009), GW2 (Song et al., 2007), *qSW5* (Shomura et al., 2008), *GW8* (Wang et al., 2012b), and *GS5* (Li et al., 2011). Although the genome assembly of weedy rice WR04-6 has been constructed (Sun et al., 2019), the progress of identifying genes from weedy rice and the functional genomics research remains hindered due to a lack of more high-quality reference genomes. Generally, the morphological characteristics of weedy rice is between wild rice (*O. rufipogon*) and cultivated rice (*O. sativa* L.) (Sun et al., 2013; Cui et al., 2016). Our previous taxonomic study showed that LM8 is a low heterozygous weedy rice germplasm. The plants are homozygous and can be inherited stably that is characterized by very small grains. To discover genes from weedy rice germplasms of LM8, we constructed a genetic map based on whole-genome sequencing of a F_2 population derived from the cross between LM8 and a cultivated rice variety. In combination with the phenotypic data of 12 important agronomic traits collected from the F_2 population, we also tried to identify some new genes from the weedy rice. Moreover, to clarify the genomic characteristics from weedy rice germplasms of LM8, we generated a high-quality genome assembly of LM8 based on the Nanopore sequencing technology and characterized the LM8 genome to reveal its evolutionary relationship, which broadens our understanding of weedy rice at the genomic level. Based on our study, we found that the combination of genetic map and genome map is critical to quickly discover candidate genes such as plant-type, panicle-type, and gain-size in weed rice.

MATERIALS AND METHODS

Plant Materials

The weedy rice LM8 was obtained from the China National Genebank. It shows erect and compact architecture similar to cultivated rice and harbors typical characteristics, such as small grain size and black hull. The cultivated rice variety Shen 08S was provided by the Anhui Academy of Agricultural Sciences. A F_2 population (1229 samples) was obtained from a cross between LM8 and Shen 08S and was planted in the experimental fields under natural growth conditions in Nanning, Guangxi Autonomous Region, China. In this study, the F2 population were collected from one F1. Fresh and healthy leaves were collected at seedling stage and stored at 80°C for subsequent genomic DNA extraction.

Population Sequencing and Genetic Map Construction

Fresh leaves of randomly selected 199 samples of the F_2 population and their parents (LM8 and Shen 08S) were used to extract genomic DNA with the cetyltrimethylammonium bromide (CTAB) method. The Illumina PE150 libraries were constructed according to the manufacturer's instructions and sequenced on an Illumina HiSeq X Ten platform. The two parental genotypes were sequenced at a higher depth (20 × coverage) to obtain 10 Gb data each, and F_2 individuals were sequenced at a lower depth (∼ 10 × coverage) to obtain 5 Gb data each. Low-quality reads were removed to obtain clean reads, which were then mapped to the LM8 genome (LM8_v1) using BWA (mem -t 4 -k 32 -M -R) (Li and Durbin, 2009). SAMtools (sort rmup) (Li et al., 2009) was used to convert and sort the mapping results and to remove PCR duplicate reads. The clean reads of each F_2 individual that passed the quality control were mapped to the reference genome (LM8_v1) for haplotype-based SNP calling. Development of polymorphic markers was performed by GATK (McKenna et al., 2010) for SNP identification and genotyping, and a total of 2,373,849 SNP

markers were obtained. Then, these SNP markers were filtered by removing abnormal bases, abnormal genotypes, incomplete coverage markers, and segregation distortion markers, and were sorted into LGs (Yang et al., 2018). After filtering, 10,739 SNP markers were cluster into 12 LGs using JoinMap v4.1 (Mapping algorithm—ML Mapping, Regression mapping—Kosambi's) (Stam, 1993).

Phenotypic Evaluation of the F_2 Population

We collected the main culm of plant individuals at 25 days after heading to measure the plant height (PH), tillering number (TN), flag leaf length (FLL), and flag leaf width (FLW) using a ruler. At maturity, the main panicles of plant individuals were harvested to measure panicle length (PL) using a ruler, and the primary branch number (PB) and secondary branch number (SB) (Ma et al., 2016) were recorded. The filled grains were used to calculate the grain length (GL), grain width (GW), grain thickness (GT), length width ratio (LWR), and thousand-grain weight (TGW) using an automatic seed analyzer with three replicates (Wanshen Detection Technology, Hangzhou, China). The analysis of variance (ANOVA) and correlations of phenotypic characteristics collected from the F_2 population were conducted in R v3.6.2 (Langfelder and Horvath, 2012).

QTL Mapping and Candidate Gene Prediction

QTL mapping was conducted using a permutation test ($n = 1{,}000$) in MapQTL6.0 with the composite interval mapping method to determine the limit of detection (LOD) value of each phenotype (Ooijen et al., 2009). Then the CIM mapping method in Win QTL Cartographer v2.5 software was used to locate the QTL position, contribution rate, and additive effect (Wang et al., 2012a). The 99% confidence interval of a QTL were determined as a candidate region, in which genes harbored non-synonymous coding mutations, premature or extended termination mutations were regarded as functional genes.

Genome Library Construction and Sequencing

Genomic DNA was extracted from the fresh leaves of LM8 using Genomic kit (13343, Qiagen, Germany). Total RNA was extracted from five different tissues (root, leaf, stem, flower, and spike) by using the TRNzol Universal Total RNA extraction Kit (DP424, Tiangen, China). The total RNA was reserve transcribed into cDNA using SMARTer PCR cDNA Synthesis Kit (634926, Takara, China). PCR was performed using PrimeSTAR GXL DNApolymerase (R050A, Takara, China). The purity, concentration, and integrity of DNA and RNA were determined using NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, United States), Qubit® 3.0 Fluorometer (Invitrogen, United States) and Agilent 2100 Bioanalyzer (Agilent technologies, United States).

A library for Illumina paired-end sequencing with an insert size of 350–500 bp was constructed and sequenced on an Illumina HiSeq X ten platform (Illumina, San Diego, CA, United States). Oxford Nanopore library preparation was conducted according to the manufacturer's instruction (13343, Qiagen, Germany) and sequenced on a PromethION platform (Oxford Nanopore Technologies, Oxford, United Kingdom). Fresh young leaves were vacuum-infiltrated with formaldehyde solution and used for cross-link action. The Hi-C library was prepared following the manufacturer's protocol and sequenced on an Illumina HiSeq X ten platform. SMRTbell library of RNA-seq was constructed from a pooled cDNA sample of five different tissue (root, leaf, stem, flower, and spike) using SMRTbell template prep kit 2.0 (100222300, Pacific Biosciences, United States) and sequenced on a PacBio Sequel sequencer (Pacific Biosciences, Menlo Park, United States) to obtain full-length transcriptome data.

Genome Assembly

The Illumina short reads were filtered using fastp v0.20.0 with default parameters (Chen et al., 2018). The abundance of 17 nt K-mers (-C -m 17 -s 400M) was used to estimate the genome size and heterozygous rate (Marçais and Kingsford, 2011; Liu et al., 2013; Koren et al., 2017). Correction of long reads generated from the Oxford Nanopore PromethION platform and *de novo* assembly were performed by *NextDenovo* v1.1.1 (read_cuoff = 2 k, seed_cutoff = 23 k, blocksize = 1 g, pa_raw_align = 20, pa_correction = 35) and *SMARTdenovo* (-e dom -J 5000 -k 17) (Loman et al., 2015; Cali et al., 2018). The Illumina short reads were mapped to the initial sequence assembly using BWA v0.7.12-r1039 with default parameters, which was then iteratively polished with three rounds of correction using NextPolish v3.0.1 (-max_depth 100 cluster_optons = -w n -l vf = {vf} -q all.q -pe smp {cpu} genome_size = auto) (Walker et al., 2014; Hu et al., 2020). Purge Haplotigs software was used to generate a contig-level assembly with only one copy of each of the contigs from heterozygous regions. The completeness of the draft genome was assessed by BUSCO v3 with the embryophyta_odb9 database (Simão et al., 2015).

Ultra-high-molecular-weight (uHMW) DNA (DNA length > 250 kb) were extracted using Bionano Prep Plant DNA Isolation Kit (80003; Bionano Genomics, United States) according to the manufacturer's instructions. uHMW DNA molecules were labeled with the DLE-1 enzyme and loaded onto a Saphyr Chip and scanned for images on a Bionano Saphyr system (Bionano Genomics, San Diego, CA, United States). The raw molecules generated were quality-controlled and filtered (molecules with a size < 150 kb were removed). An optical map was generated using Bionano Solve package v3.4. The generated optical map was used to construct scaffolds using the Hybrid Scaffold pipeline of Bionano Solve package v3.4 (CL.py -d -U -N 6 -y -i 3 -F 1 -a opt Arguments_non-haplotype_noES_noCut_saphyr.xml) and Bionano Access v1.5.2 (Bionano Genomics, San Diego, CA, United States) with a more stringent (1e-13) merging *p*-value threshold (Xiao et al., 2007; Reisner et al., 2010; Mostovoy et al., 2016). The Hi-C raw reads were filtered by fastp v0.12.6 with default parameters and then mapped to the scaffolds with Bowtie2 (Langmead and Salzberg, 2012; Chen et al., 2018). We used Lachesis (ligating adjacent chromatin enables scaffolding *in situ*) to cluster, order, and

anchor scaffolds onto the chromosomes (Burton et al., 2013; Dudchenko et al., 2017).

Annotation of Genome

The repeat sequences and elements were annotated by a combination of *de novo* and homology-based methods. LTR_FINDER (Haas et al., 2008) and RepeatModeler (Haas et al., 2003) were used to generate a dataset of repetitive sequences with default parameters. This dataset was BLAST against the Plant Genome and Systems Biology (PGSB) repeat element database to classify the repeats (Spannagl et al., 2016), and then RepeatMasker was employed to annotate these repeats based on the Repbase database (Bao et al., 2015). Further, tandem repeats finder software was used to identify tandem repeats (Benson, 1999).

The protein-coding genes of the LM8 genome were predicted through a comprehensive strategy that combined results obtained from *de novo*, homology-based, and transcriptome-based predictions. Augustus was used for *de novo* prediction with Hidden Markov Model (Stanke et al., 2008). Homologous proteins from six plant genomes (*Arabidopsis thaliana*, *O. sativa*, *Zea mays*, *Hordeum vulgare*, *Physcomitrella patens*, and *Triticum aestivum*) were downloaded from Ensembl plants[1] and used for homology-based prediction by GeMoMa (Jens et al., 2016). The non-redundant full-length transcripts obtained from the PacBio Sequel platform were aligned to the LM8 genome assembly for transcriptome-based prediction using PASA (Haas et al., 2003).

Gene structures were determined based on a combination of results from the three prediction methods using EvidenceModeler (Haas et al., 2008). Functional annotation of protein-coding genes was achieved by BLASTP searches against the Swiss-Prot database (Stanke and Waack, 2003). Protein domains were annotated by searching against the InterPro database using InterProScan (Zdobnov and Apweiler, 2001; Hunter et al., 2009). Non-coding RNA genes, including miRNA, snRNA, and rRNA genes were predicted according to the Rfam database, while tRNA genes were identified using tRNAscan-SE (Lowe and Eddy, 1997; Griffiths-Jones et al., 2005). The completeness of the predicted gene set was assessed by BUSCO v3 with the embryophyta_odb9 database (Benson, 1999).

Collinearity Analysis

Protein sequences of LM8, *japonica* var. Nipponbare, and *indica* var. R498 were aligned by BLASTP v2.6.0 with default settings. Syntenic gene blocks within the genome were detected by MCScanX (Wang et al., 2012c) and visualized using the jcvi python module.

Identification of Gene Families

Gene family identification was performed across LM8 (*O. sativa* f. *spontanea*), *O. aus* (AUS), 5 cultivated rice varieties, and 11 wild rice species. The 5 cultivated rice varieties included *O. sativa* ssp. *indica* (IND), *O. sativa* ssp. *japonica* (JAP), *O. sativa* ssp. *indica* var. Minghui63 (MH63), *O. sativa* ssp. *indica* var. Zhenshan97 (ZS97), *O. sativa* ssp. *indica* var. Shuhui498 (R498). The 11 wild rice species consisted of *O. glaberrima* (GLA), *O. barthii* (BAR), *O. glumaepatula* (GLU), *O. meridionalis* (MER), *O. rufipogon* (RUF), *O. nivara* (NIV), *O. longistaminata* (LON), *O. punctata* (PUN), *O. brachyantha* (BRA), *O. rufipogon* var. JX-6 (JX-6), and *O. rufipogon* var. Z59 (Z59). PUN and BRA belong to the BB and FF genomes, respectively, while the others belong to the AA genome. Across all species, the longest transcript of each gene was used in further analyses. Orthologous and paralogous gene clusters were identified using BLASTP (-e 1e-5 -F F). Clustering analysis of protein sequences from the 18 *Oryza* genomes was conducted with OrthoMCL (Li et al., 2003).

Phylogenetic Analysis

Multiple sequence alignments of the protein-coding sequences of the 4,241 single-copy orthologous genes obtained from the above analysis these protein sequences were performed by MAFFT (Katoh and Standley, 2013). Phylogenetic relationships were resolved using RAxML (-m GTRGAMMA -p 12345 -T 8 -f b -t -z) among these 18 *Oryza* genomes with all single-copy genes concatenated into an ultra-long aligned sequence, where *O. brachyantha* was designated as an outgroup (Stamatakis, 2014). Divergence times were estimated by MCMCtree (Puttick, 2019) with parameters of "RootAge $\leq$ 0.21, rgene gamma = 23.52254, burnin = 100,000, sampfreq = 100, nsample = 50,000, model = 7" in the PAML package (Nikolau et al., 2003) based on a known divergence time ($\sim$ 0.4 Mya) between *O. nivara* and *O. rufipogon*.

Expansion and Contraction of Gene Families

A random birth-and-death model was used to estimate changes in gene families between the ancestor and each species using CAFE with conditional likelihoods as the test statistics (-p 0.05 -t 10 -r 10000 lambda -s) (De Bie et al., 2006). A probabilistic graphical model (PGM) was used to calculate the probability of transitions in each gene family, and then all the gene families were classified into three types (expanded, contracted, and unchanged). Finally, GO enrichment was performed for further functional analysis of the expanded genes.

Positive Selection Analysis

All orthologous genes identified in the LM8 genome were tested for positive selection. The phylogenetic tree generated by RAxML was used as the input, and the branch-site test was conducted with CodeML (model = 2, NSsites = 2, fix_omega = 0, fix_omega = 1, omega = 1) in the PAML package (Yang, 2007). Genes under positive selection were determined based on the likelihood ratio test ($P < 0.01$).

RESULTS

Genetic Map Construction and QTL Analysis With a F_2 Population

To further understand the mechanism of LM8 genome variation in its special grain formation, a F_2 population was generated from the cross between LM8 and a cultivated rice variety Shen 08S.

[1] http://plants.ensembl.org

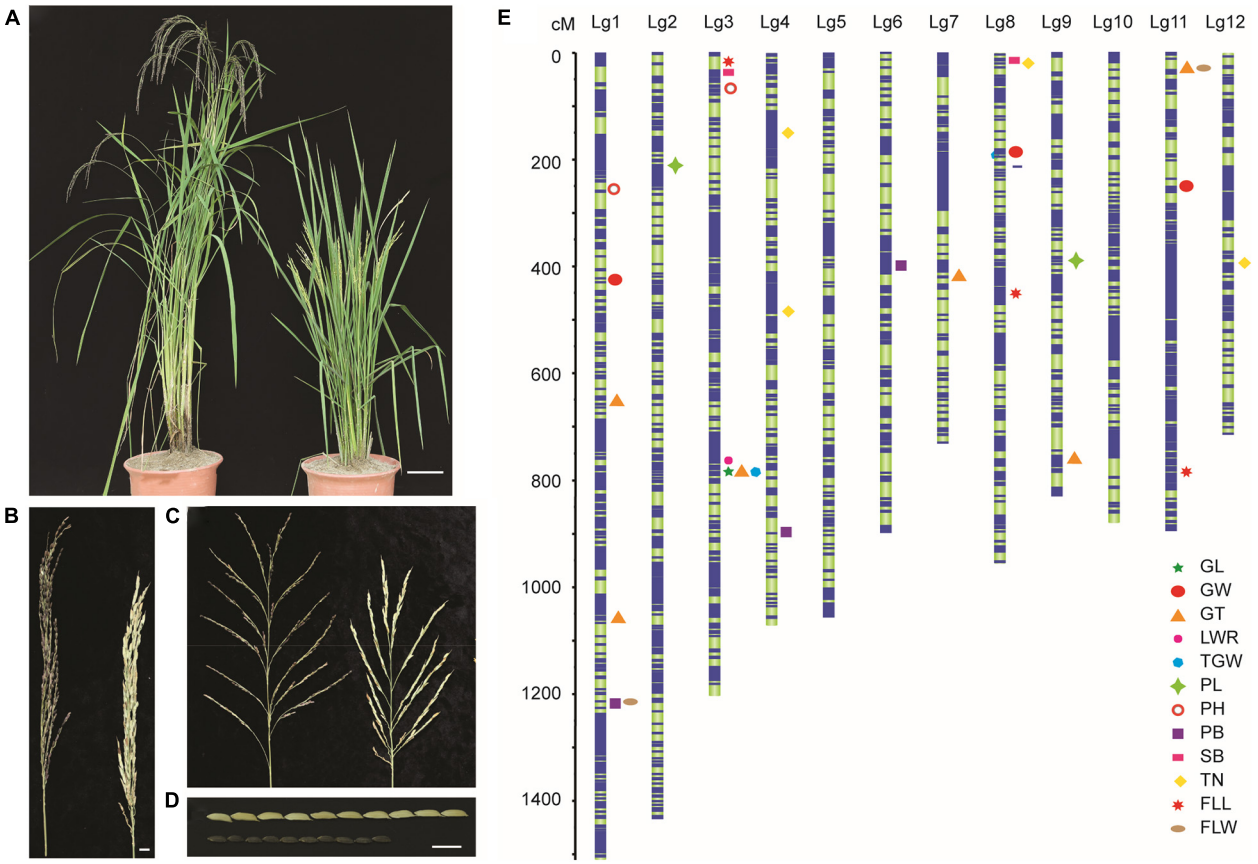


FIGURE 1 | Agronomic characteristics of LM8 and Shen 08S (parents) and the genetic map constructed based on the F_2 population. **(A)** Plant height of LM8 (right) and Shen 08S (left). Bar = 10 cm. **(B)** Panicle length of LM8 (right) and Shen 08S (left). Bar = 0.5 cm. **(C)** Branches of LM8 (right) and Shen 08S (left). Bar = 2 cm. **(D)** Grain size of LM8 (bottom) and Shen 08S (above). Bar = 1 cm. **(E)** Genomic locations of 12 agronomic trait associated QTLs are illustrated on the 12 linkage groups. PH, plant height; TN, tillering number; FLL, flag leaf length; FLW, flag leaf width; PL, panicle length; PB, primary branch number; SB, secondary branch number; GL, grain length; GW, grain width; GT, grain thickness; LWR, length to width ratio; TGW, thousand-grain weight.

The two parents, LM8 and Shen 08S, showed obvious differences in plant height, panicle length, and grain size (**Figure 1** and **Supplementary Table 1**). We sequenced the genome of F_2 individuals as well as that of the two parents. A total of 10,739 high-quality SNPs were obtained and used to generate a genetic map. The total genetic distance of the constructed genetic map was 12,171.13 cM, and the average genetic distance between two SNPs was 1.13 cM (**Figure 1**). The SNPs were distributed throughout the 12 linkage groups (LGs) with the highest SNP number (1,754) occurring on LG1 (1,510.48 cM total size) and the lowest (523) on LG12 (713.79 cM total size). Collinearity analysis showed that the genetic map had strong collinearity (99.69%) with the reference genome sequence (**Supplementary Figure 1**), and the sources of most segments in F_2 individuals were consistent according to the monomer source analysis. These results suggest that the constructed genetic map is of high-quality and suitable for further analyses.

Besides, combining the phenotypic data (**Figure 2**) obtained from the F_2 population and the genetic map, we identified 31 quantitative trait loci (QTLs) with 607 genes related to 4 plant-type traits, 3 panicle-type traits, and 5 grain-size traits (**Figure 1**). Eight of the QTLs explaining more than 17% of the phenotypical variation were identified as major QTLs, which were located at 788.3–789.4 cM on chromosome 3 (chr3), 34.4–37.5 cM on chr11, 782.9–786.6 cM on chr3, 787.6–788.2 cM on chr3, 244.6–253 cM on ch11, 204.9–217.6 cM on chr2, 11.4–17.3 cM on chr8, and 33.6–38 cM on chr11 (**Supplementary Table 2** and **Supplementary Figure 2**). Fourteen QTLs were identified to be associated with grain-size traits, including 1 for grain length (GL), 3 for grain width (GW), 6 for grain thickness (GT), 1 for length width ratio (LWR), and 3 for thousand-grain weight (TGW). These results would help in further detecting the genes from the weedy rice LM8.

Identification of Candidate Genes Related to Grain Length

LM8 has evolved to form extremely small grains that may develop new elite rice varieties to study grain shape or yield related traits. Therefore, using LM8 as the material to discover genes related to grain size is practical to enrich rice resources. We conducted a correlation analysis among the grain-size traits QTLs, including

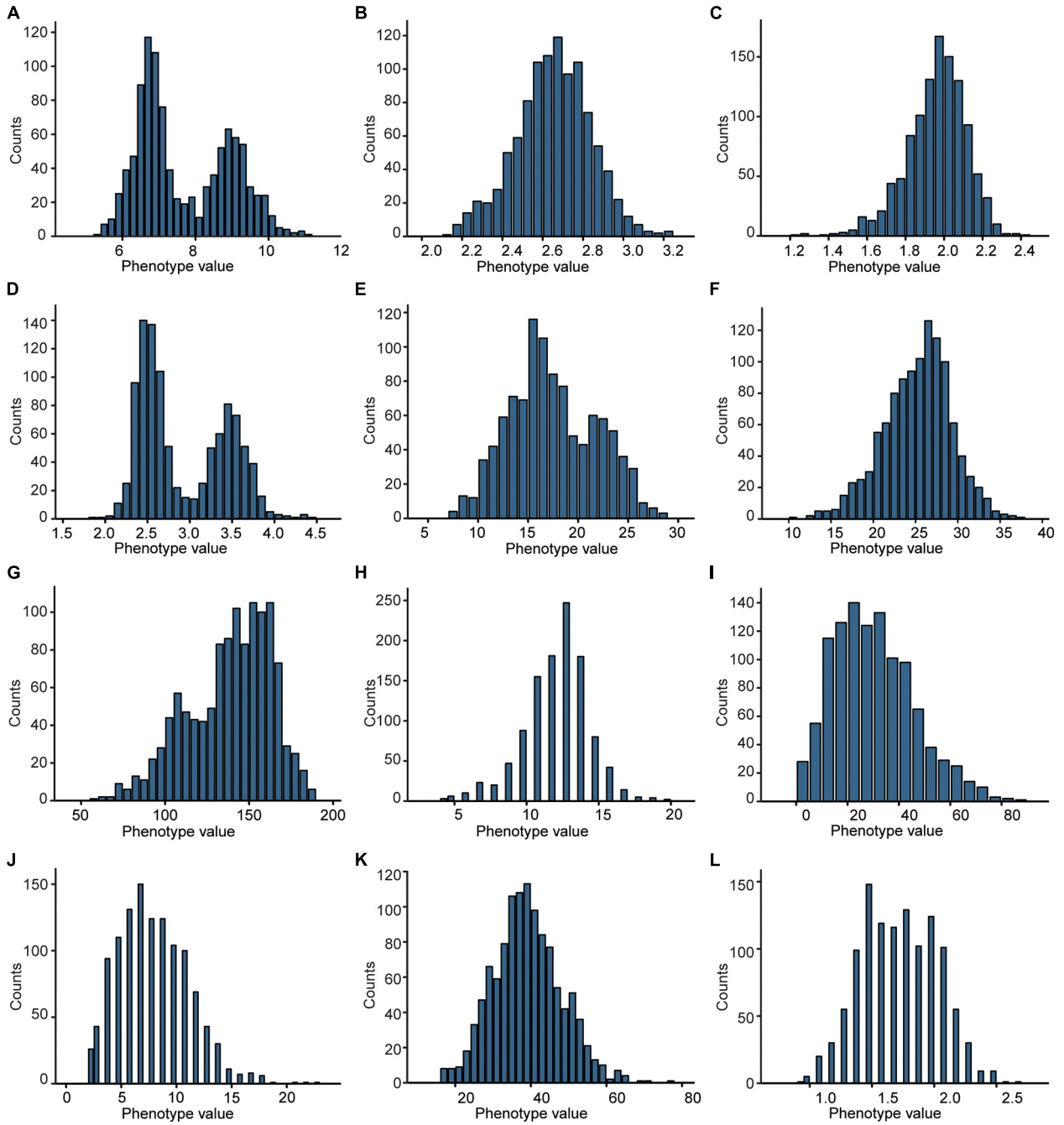


FIGURE 2 | Frequency distribution of the 12 agronomic traits in the F_2 population. **(A)** GL, grain length. **(B)** GW, grain width. **(C)** GT, grain thickness. **(D)** LWR, length to width ratio. **(E)** TGW, thousand-grain weight. **(F)** PL, panicle length. **(G)** PH, plant height. **(H)** PB, primary branch number. **(I)** SB, secondary branch number. **(J)** TN, tillering number. **(K)** FLL, flag leaf length. **(L)** FLW, flag leaf width.

grain length, grain width, grain thickness, length to width ratio, and thousand-grain weight. Significant positive correlations ($P < 0.05$) were observed among grain length, length to width ratio, and thousand-grain weight, indicating that grain length has significant impact on grain size (**Supplementary Table 3** and **Supplementary Figure 3**). One QTL related to grain length was located at 788.3–789.4 cM on chr3, corresponding to a 60-kb interval harboring seven putative genes, which included 3 RAPdb annotated genes (*ORUFILM03g000091*, *ORUFILM03g000095*, *ORUFILM03g000096*) and 4 unknown function annotations (*ORUFILM03g000090*, *ORUFILM03g000092*, *ORUFILM03g000093*, *ORUFILM03g000094*) were important candidate genes controlling grain length (**Figure 3** and **Supplementary Table 4**).

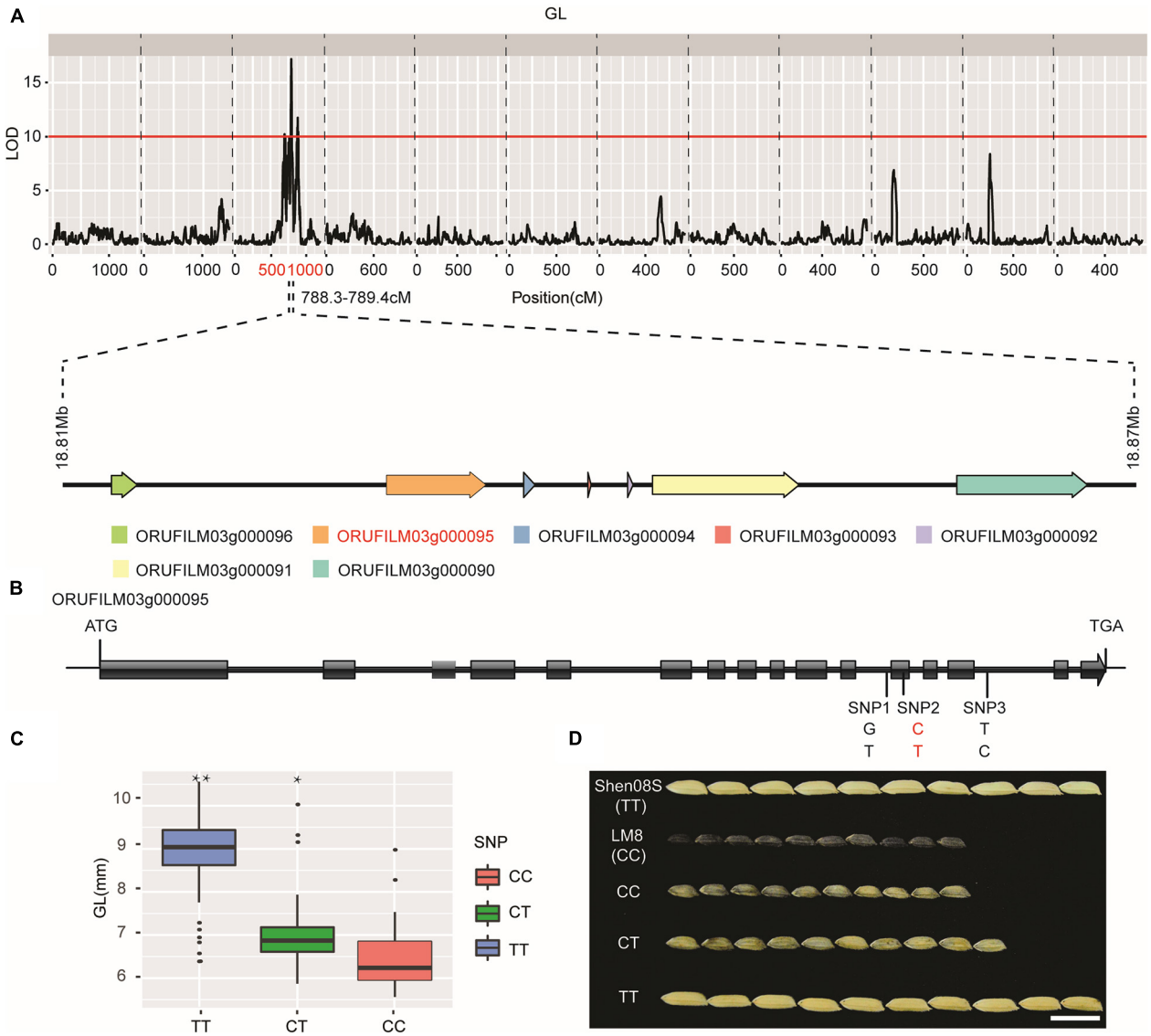


FIGURE 3 | QTL for grain length and prediction of candidate genes. **(A)** QTL mapping results for grain length and seven putative genes (*ORUFILM03g000096*, *ORUFILM03g000095*, *ORUFILM03g000094*, *ORUFILM03g000093*, *ORUFILM03g000092*, *ORUFILM03g000091*, and *ORUFILM03g000090*) within the region. **(B)** Gene structure of *ORUFILM03g000095* (top), **(C)** grain length (GL) distribution in each genotype of *ORUFILM03g000095*, **(D)** corresponding GL phenotypes in F_2 individuals (lower right). Asterisks is correlation between genotype and phenotype, $^*P < 0.05$. $^{**}P < 0.01$.

OsCLG1 (Yang et al., 2021) mediate ubiquitin ligase to regulate grain length. Therefore, the candidate genes among seven candidate genes, *ORUFILM03g000095* is a homologous gene to *Os03g0427900* of *Nipponbare* and belongs to the U-box protein gene family, in which a U-box domain acts as a ubiquitin ligase to participate in protein degradation during the cell cycle and morphological development (Sharma and Taganna, 2020; Yang et al., 2021). To further investigate the molecular basis of the small grain phenotype in LM8, we analyzed the sequence of *ORUFILM03g000095* gene from LM8, Shen 08S, and their progenies and revealed a C-T SNP site, located in the 12th exon 5,339 bp downstream of the ATG start site (**Figure 3**). Grain length in the F_2 individuals of LM8 and Shen 08S displayed a clear pattern with an order of TT > CT > CC ($P < 0.01$; **Figure 3**). *ORUFILM03g000095* genotypes were significantly correlated to the grain length variation, suggesting that this locus plays an important role in grain size regulation. Our results suggest that *ORUFILM03g000095* are possible candidate genes controlling grain length. However, the underlying mechanisms of how this gene regulate grain formation remain elusive and need to be further explored.

Genome Assembly and Annotation

There are major differences between the morphology of weedy rice and cultivated rice (*O. sativa* L.). The current research on cultivated rice is relatively clear, but the research on

weedy rice does not yet have a reference genome with high assembly quality. To clarify the genome characteristics of the F2 population parent (weed rice LM8), we assembled a high-quality genome. Before assembly, *SOAPdenovo* was used for pre-assembly. K-mer analysis (k = 17) estimated its genome size to be around 362.7 Mb, with a moderate heterozygous rate of 0.20% (**Supplementary Figure 4**). However, the completeness and quality of the assembly are not ideal if the genome is assembled using the second-generation sequencing data alone. Thus, the LM8 genome was sequenced and assembled by applying a combination of diverse technologies, including Oxford Nanopore long-read sequencing, Illumina short-read sequencing, Bionano optical mapping, and Hi-C technology (**Supplementary Table 5** and **Supplementary Figure 5**). A total of 77.2 Gb raw data (sequencing depth 100x) were collected from Oxford Nanopore long-read sequencing, which were then self-corrected, filtered, and polished to generate the final dataset (57.3 Gb) for genome assembly (**Table 1** and **Supplementary Figure 6**). The contig-level assembly (LM8_contig) comprised 375.3 Mb, with a contig N50 of 17.9 Mb (**Table 1** and **Supplementary Table 6**). Approximately 98.1% ubiquitous genes in embryophyte were detected by the Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (**Supplementary Table 7**), indicating that the assembled contig was of high-completeness.

TABLE 1 | Summary of the sequencing, assembly, and annotation of the LM8 genome.

Stat type	Number
Assembled genome size (Gb)	77.2
Contig N50 (Mb)	17.9
Scaffold N50 (Mb)	30.5
Longest scaffold (Mb)	31.3
Anchored to chromosome (Mb)	375.8
Number of predicted protein-coding genes	36,561
Average gene length (bp)	3545.1
Average CDS length (bp)	1129.6
Average exons number per gene	4.4
Average exon length (bp)	255
Average intron length (bp)	705
Number of rRNAs	81
Number of snRNAs	772
Number of miRNAs	2,551

Next, using 476.2 Gb of molecules (> 150 kb) collected from Bionano Saphyr system, we generated an optical map for the LM8 genome, with a total size of 370.3 Mb and an N50 of 24.2 Mb. With the aid of this optical map, we further assembled LM8_contig into scaffolds (LM8_scaffold),

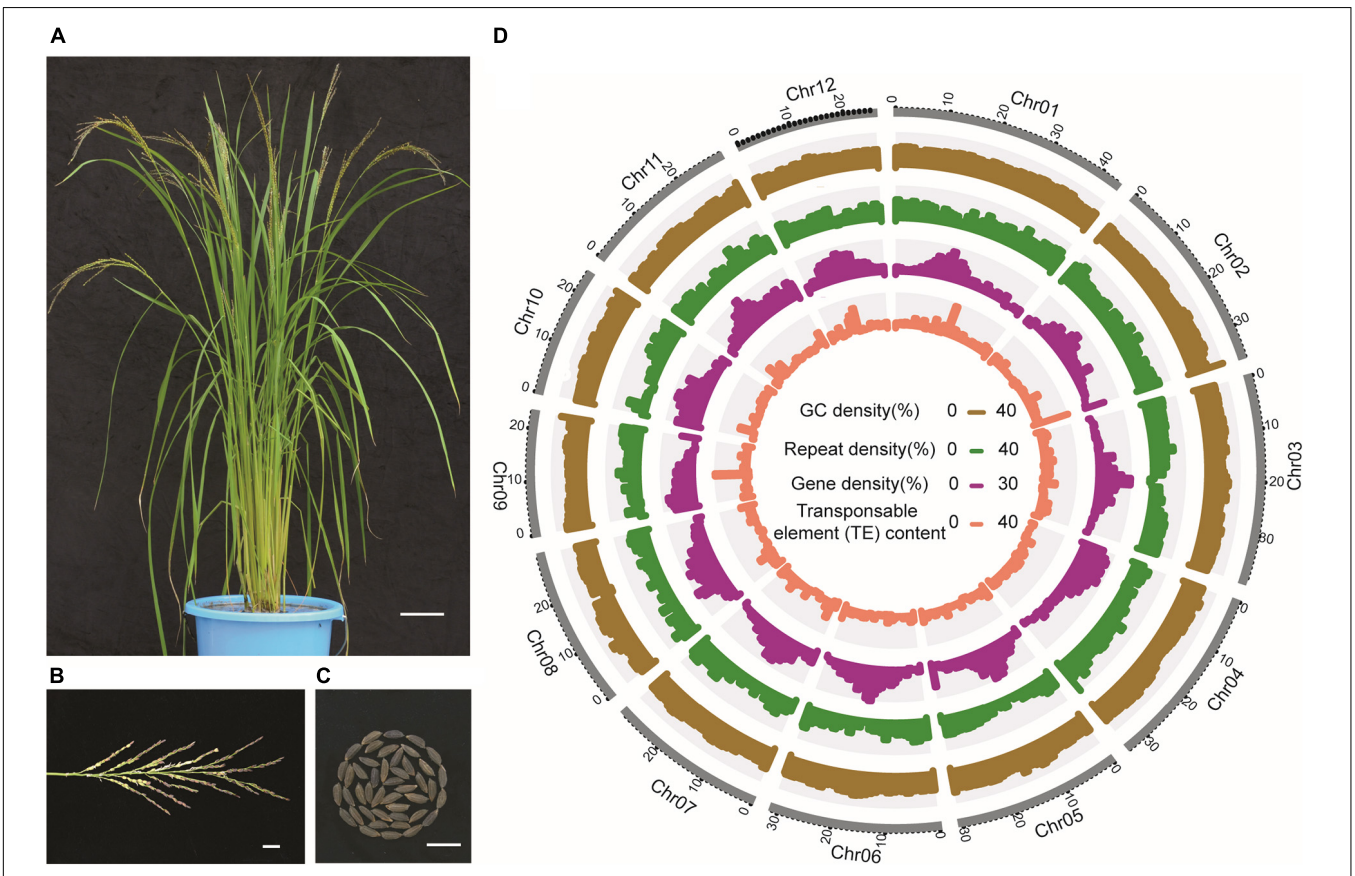


FIGURE 4 | The morphology and genome features of LM8. **(A)** Whole plants, bar = 10 cm. **(B)** Panicles, bar = 2 cm. **(C)** Grains, bar = 1 cm. **(D)** Genome features of LM8. Circles from inside to outside are transposable element (TE) content, repeat density, gene density, and GC density.

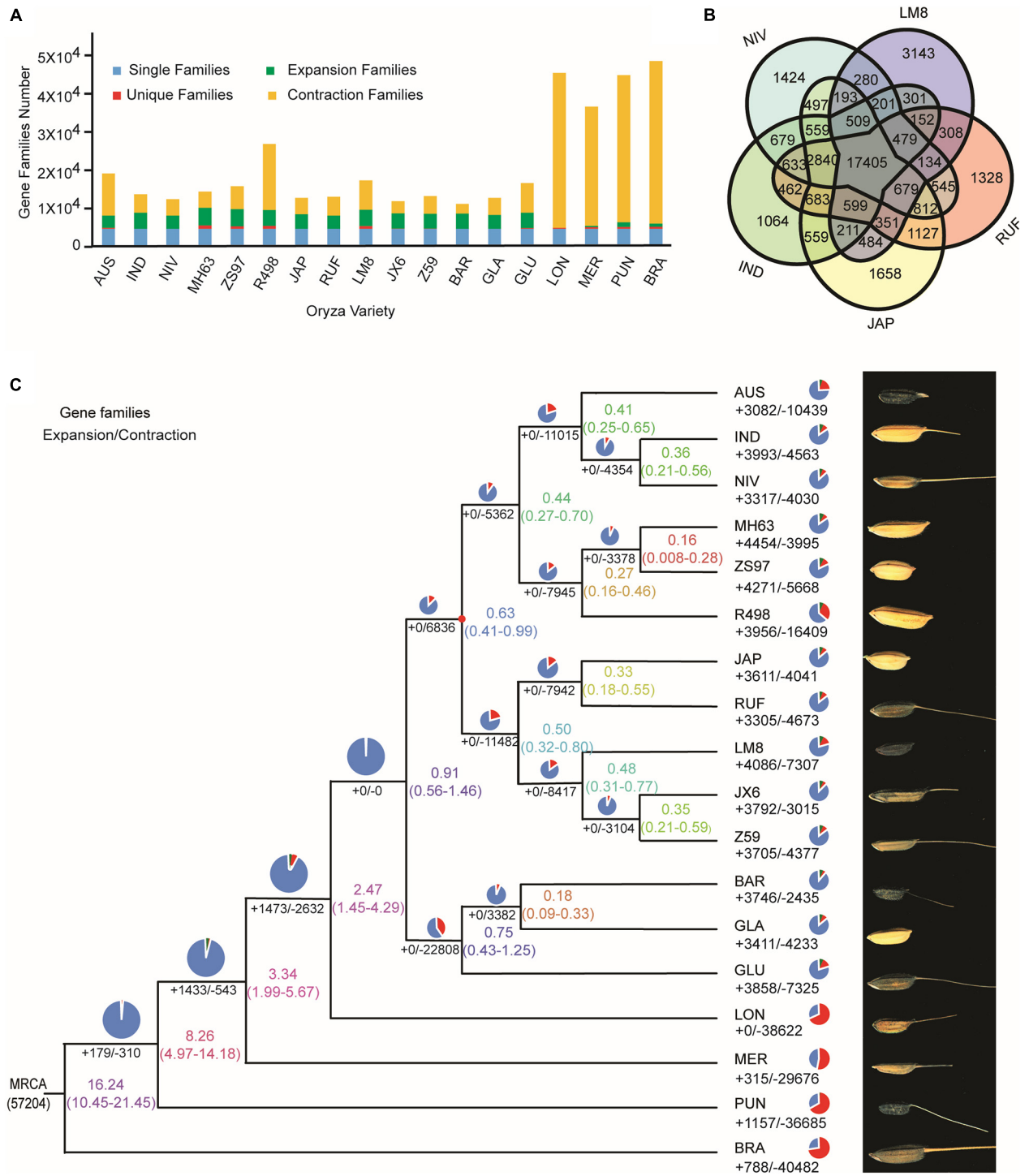


FIGURE 5 | Comparative genomics analyses of LM8 with other *Oryza* genomes. **(A)** Statistics of gene families in 18 *Oryza* genomes. **(B)** Core and dispensable genes from five reference genomes. The numbers in the species section and overlapping section indicate the numbers of specific and shared gene families, respectively. IND, *O. sativa* ssp. *indica*. JAP, *O. sativa* ssp. *japonica*. RUF, *O. rufipogon*. NIV, *O. nivara*. and LM8, *O. sativa* f. *spontanea*. **(C)** Phylogenetic relationships and grain phenotypes of LM8 and other *Oryza* genomes. Pie charts represent total gene families, consisting of contracted gene families (red), expanded gene families (green), and unchanged gene families (blue). The numbers of genes in expanded (+) and contracted (–) gene families in each rice variety are shown with the rice variety name farthest to the right. The lineage divergence times are indicated on the nodes and nodes marked in red are known fossil time points. *O. brachyantha* was used as the outgroup. MRCA, most recent common ancestor. AUS, *O. aus*. IND, *O. sativa* ssp. *indica*. JAP, *O. sativa* ssp. *japonica*. GLA, *O. glaberrima*. BAR, *O. barthii*. GLU, *O. glumaepatula*. MER, *O. meridionalis*. RUF, *O. rufipogon*. NIV, *O. nivara*. LON, *O. longistaminata*. PUN, *O. punctata*. BRA, *O. brachyantha*. JX-6, *O. rufipogon* var. JX-6. Z59, *O. rufipogon* var. Z59. MH63, *O. sativa* ssp. *indica* var. Minghui63. ZS97, *O. sativa* ssp. *indica* var. Zhenshan97. R498, *O. sativa* ssp. *indica* var. Shuihui498. and LM8, *O. sativa* f. *spontanea*.

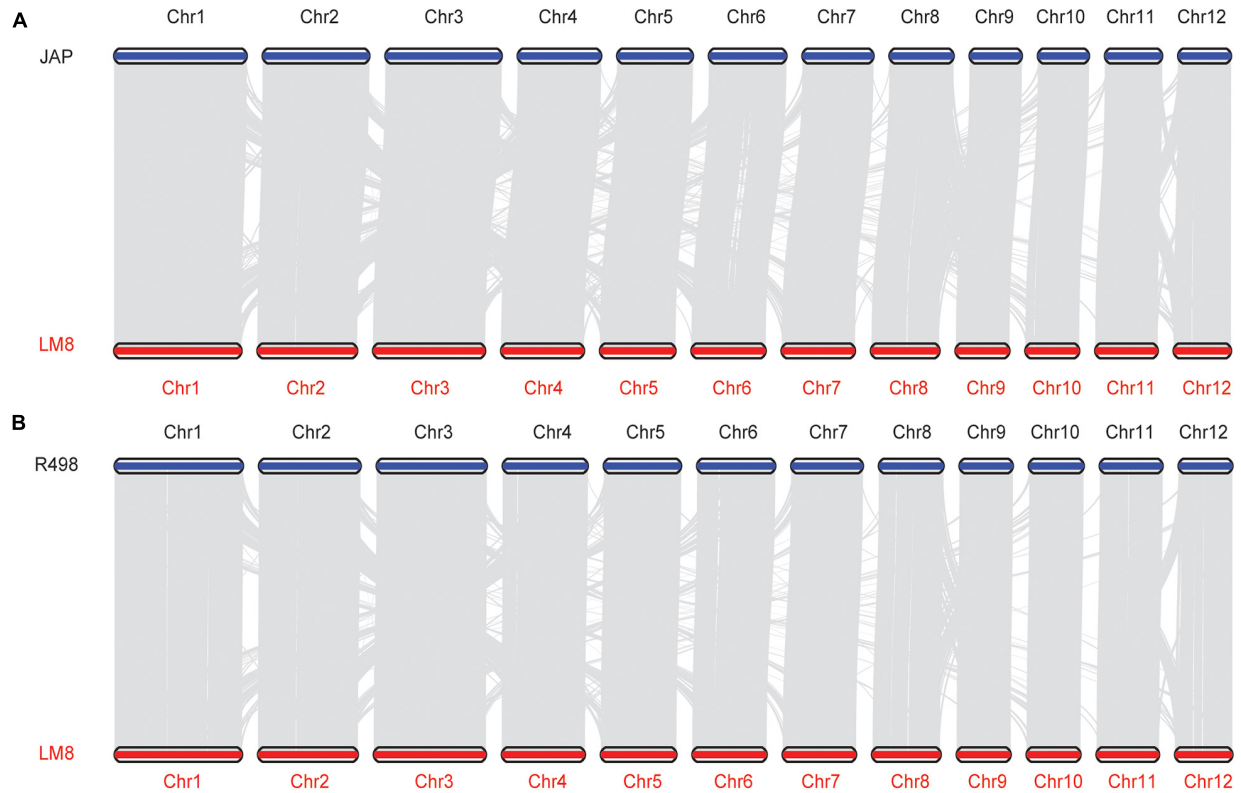


FIGURE 6 | Collinearity between LM8 and two cultivated rice (JAP and R498) genomes. **(A)** Collinearity between JAP and LM8. **(B)** Collinearity between R498 and LM8.

with a total size of 375.8 Mb and a scaffold N50 of 24.1 Mb (**Supplementary Table 6**). After applying high-throughput chromosome conformation capture (Hi-C) data to orient, order, and phase these scaffolds, a total of 375.3 Mb sequences (99.85%; **Supplementary Table 8**) were anchored onto the 12 chromosomes and the final chromosome-level genome assembly (LM8_v1) was obtained. The Hi-C heatmap separated different chromosomes and showed that the interaction intensity in the diagonal-position was higher than that in the off-diagonal-position (**Supplementary Figure 7**). BUSCO analysis showed that 97.9% of the core embryophyte genes were complete in the LM8 genome assembly (**Supplementary Table 7**). In addition, 87.1% (31,810) of the predicted genes were expressed according to the transcriptome data. The above results suggest that the LM8 genome assembly is of high-quality and -completeness.

Repeat annotation results showed that 47.72% of the LM8 genome is composed of repetitive sequences, including 26.87% retrotransposons and 20.85% DNA transposons. About 94.08% of retrotransposons are long terminal repeats (LTRs), accounting for 25.28% of the genome. The two most frequent types of LTRs are *Copia* and *Gypsy*, accounting for 2.99 and 19.62%, respectively (**Figure 4** and **Supplementary Table 9**). Besides, through a comprehensive strategy combining results obtained from *de novo*, homology-based, and transcriptome-based prediction, 36,561 protein-coding genes were annotated in the LM8 genome. These protein-coding genes have an average length of 3,545.1 bp, an average coding sequence length of 1,129.6 bp, an average exon length of 255.2 bp, an average intron length of 705.1 bp, and an average exon number per gene of 4.4 (**Table 1**). Among these annotated genes, 34,773 (95.91%) were functionally annotated by at least one of the Swiss-Prot, KEGG, and InterPro databases (**Supplementary Table 10**). In addition, homology-based annotation of non-coding RNAs (ncRNAs) predicted 2,551 microRNAs (miRNAs), 81 ribosomal RNAs (rRNAs), and 772 small nuclear RNAs (snRNAs; **Supplementary Table 11**).

Comparative Analysis

To reveal the evolutionary relationship of the weedy rice LM8, 4,241 single-copy orthologous genes of LM8 and those from other 17 *Oryza* genomes were used to construct a phylogenetic tree by the maximum-likelihood (ML) method (**Figure 5** and **Supplementary Table 12** and **Supplementary Figure 8**). The phylogenetic tree demonstrated that LM8 diverged from the ancestor *O. rufipogon* ∼ 0.32 million years ago (Mya; **Figure 5**) and was clustered into a group with *japonica*, indicating LM8 is more closely related to *japonica* compared to *indica*. Additionally, genome collinearity analyses conducted between LM8 and two cultivated rice varieties revealed that the LM8 genome had more collinear genes with *japonica* var. Nipponbare (47,439/78,939; 60.1%) than *indica* var. R498 (34,750/74,110; 46.89%; **Figure 6** and **Supplementary Figure 9**). Collectively, we speculate that LM8 belongs to *japonica*-type weedy rice.

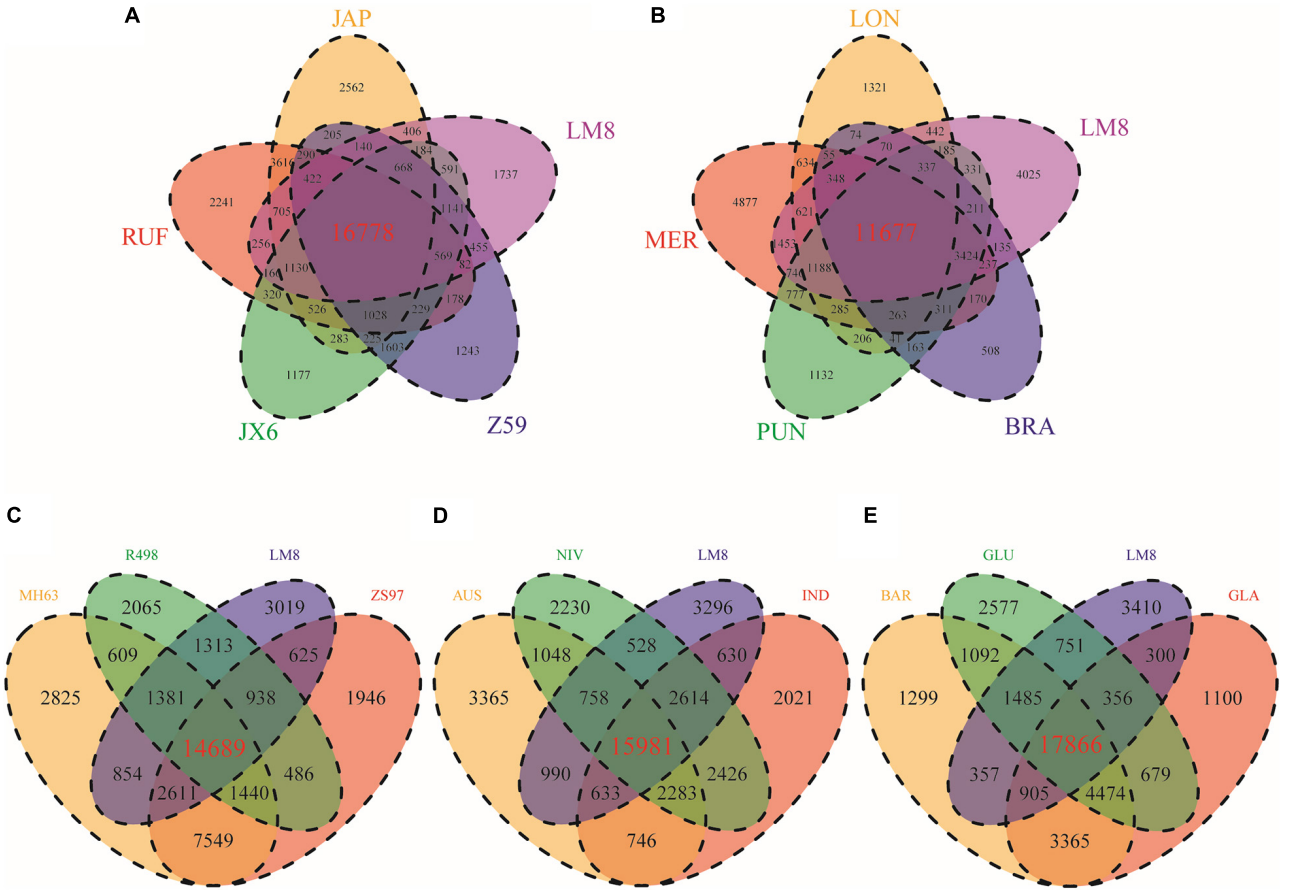


FIGURE 7 | Venn diagrams showing specific and shared gene families between LM8 and other *Oryza* genomes in the same cluster of the evolution tree. **(A)** JAP, RUF, JX6, Z59 and LM8. **(B)** LON, MER, PUN, BRA and LM8. **(C)** MH63, R498, ZS97 and LM8. **(D)** AUS, NIV, IND and LM8. **(E)** BAR, GLU, GLA, and LM8. AUS, *O. aus*. IND, *O. sativa* ssp. *indica*. JAP, *O. sativa* ssp. *japonica*. GLA, *O. glaberrima*. BAR, *O. barthii*. GLU, *O. glumaepatula*. MER, *O. meridionalis*. RUF, *O. rufipogon*. NIV, *O. nivara*. LON, *O. longistaminata*. PUN, *O. punctata*. BRA, *O. brachyantha*. JX-6, *O. rufipogon* var. JX-6. Z59, *O. rufipogon* var. Z59. MH63, *O. sativa* ssp. *indica* var. Minghui63. ZS97, *O. sativa* ssp. *indica* var. Zhenshan97. R498, *O. sativa* ssp. *indica* var. Shuihui498. and LM8, *O. sativa* f. *spontanea*.

By comparing LM8 with four other rice species including *O. nivara* (NIV), *O. sativa* ssp. *indica* (IND), *O. sativa* ssp. *japonica* (JAP), and *O. rufipogon* (RUF), we found that 68.4% (17,403/25,430) of the gene families in LM8 were shared among all five species, while approximately 12.4% (3,143/25,430) were specific to LM8 (**Figure 5**). The closer the relationship indicated by the phylogenetic tree, the more the shared gene families (**Figure 7**). Among the 18 *Oryza* genomes, 2,875 unclustered genes and 672 unique genes were observed in the LM8 genome (**Supplementary Table 12** and **Supplementary Figure 8**). The proteins encoded by these unique genes related to the formation of grain length, heading date and tillering number including serine/threonine-protein phosphatases (*ORUFILM03g000947*), photosystem II reaction center proteins (*ORUFILM08g001423*), and zinc finger MYM-type proteins (*ORUFILM03g000136*).

Gene Family Analysis

Gene family expansion/contraction has been shown to be associated with domestication and ecological adaptation (Peng et al., 2019; Zeng et al., 2019). To characterize the LM8 genome, a genome-wide comparative genomics analysis was performed among 18 *Oryza* genomes (**Supplementary Table 13**). We assigned 36,561 LM8 genes to 25,430 gene families (**Table 1** and **Supplementary Table 14**). Relative to the common ancestor of rice (*O. rufipogon*), 16.06% (4,086/25,430) expansion and 28.73% (7307/25,430) contracted gene families were observed (**Figure 5** and **Supplementary Table 14**). The expansion gene families included 12793 expansion genes, of which 213 QTL mapping genes belonged to the expanded gene family. In the expanded gene families, Gene Ontology (GO) enrichment analysis revealed 295 GO terms involving biological process (BP), cellular component (CC), and molecular function (MP). Sixty-seven pathways were significantly enriched, including carbohydrate metabolic process, signal transduction, and cell growth (**Figure 8**). The significantly enriched genes may contribute to the adaptability of LM8 to complex environments during evolution. Meanwhile, we found 57 genes among the QTL mapping were detected by GO enrichment

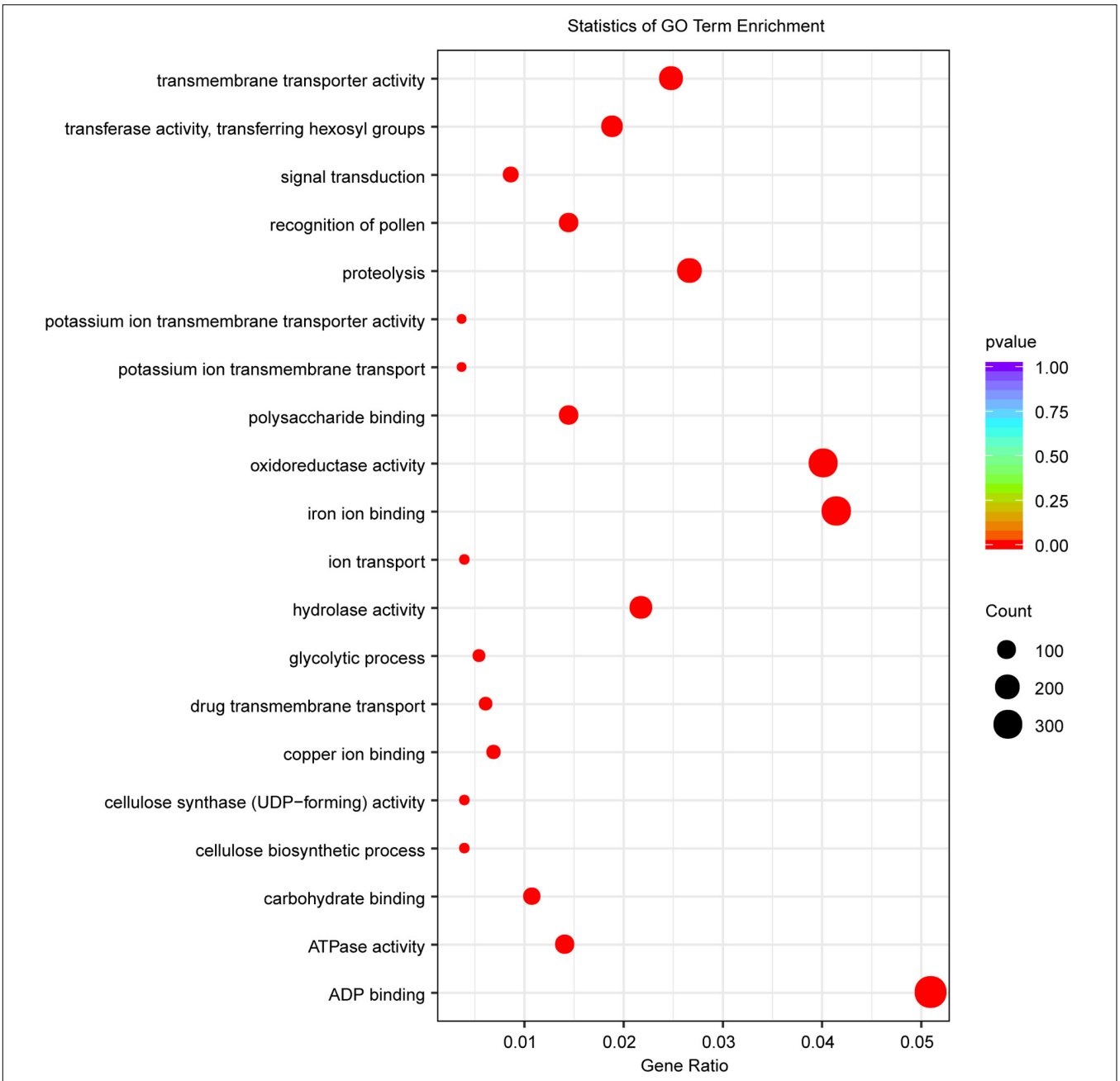


FIGURE 8 | Statistics of the GO enrichment analysis of the expanded gene families. The *x*-axis represents the percentage of enriched genes to the total annotated genes. The *y*-axis indicates the entry of each enrichment category. The size of the dots corresponds to the number of enriched genes, and the color panel on the right indicates the *q*-value. The lower the value, the more significant it is.

and enriched into 20 pathways including catalytic activity, proteolysis, and transmembrane transport protein activity (**Supplementary Table 15**). A total of 168 positive selection genes (PSGs) were identified and annotated to be auxin response proteins (e.g., *ORUFILM02g003288*), cell division control proteins (e.g., *ORUFILM01g004046*), and ubiquitin-protein ligase E3 UPL4 (e.g., *ORUFILM05g002772*), which may participate in the regulation of grain growth process and grain formation (Luo et al., 2013; Basunia et al., 2021). Nevertheless, whether these PSGs can explain the difference in the grain size need to be further explored.

DISCUSSION

With the development of sequencing techniques and corresponding analysis approaches, the sequencing speed and quality have greatly improved, while the cost has decreased

tremendously, allowing a growing number of genomes to be sequenced and applied to related studies. The combination of a specific chromosome-level genome assembly and a high-density genetic map has been verified to be effective to map QTLs or locate genes associated with important agronomic traits (Luo et al., 2020) and has been widely applied to various important crops including cotton (Wang F. et al., 2020), peanut (Agarwal et al., 2018), *Cucumis melo* (Hu et al., 2018), pear (Li et al., 2019). In rice, Li et al. (2018) constructed a high-density genetic map through performing whole-genome resequencing and identified a candidate gene (*DEP1*) in determining panicle length. Later, Sun et al. (2019) constructed a genetic map and located a region on chr1 contributing to seed shattering, awn length, and plant height. In this study, we generated a chromosome-level genome assembly and constructed a high-density genetic map with the help of high-throughput sequencing approaches, we identified *ORUFILM03g000095* gene on chr3 that may regulate grain length (**Figure 3**). We have analyzed the candidate gene based on 3K genome data which is important research in rice genomics research (Wang et al., 2018; Wang C. et al., 2020), but the same haplotype as LM8 was not found in 3K data, so we did not further analyze it (**Supplementary Table 16**). This study would not only lay a foundation for rapid discovery of genes from weedy rice but also broaden the understanding of weedy rice utilization on rice genetic improvement. Large number of candidate genes were obtained in this study and those excellent gene could improve the breeding value of cultivated rice. Next step studying of the function of the candidate gene can use gene knockout, mutation analysis, overexpression analysis, genetic complementation, and other experiments to further verify whether the candidate gene can be used to improve cultivated rice.

The *Oryza* genus is generally believed to include 22 wild and 2 cultivated rice species based on morphological characteristics (Jacquemin et al., 2014). Asian cultivated rice (*O. sativa* L.), an important staple crop, is widely planted around the world and has formed extremely rich genetic diversity during the long evolutionary process. In *O. sativa*, the two subspecies (i.e., *indica* and *japonica*) differ in morphology, anatomical structure, physiological and biochemical characteristics, and genome sequence, and their origins remain controversial (Shinobu et al., 2002; Vaughan et al., 2007). The single-origin theory believes that *indica* and *japonica* both derived from *O. rufipogon* and diverged during the long-term domestication and artificial selection (Chang, 1976; Zhu and Ge, 2005). By contrast, the multi-origin theory believes that *indica* originated from *O. nivara* in eastern India, while *japonica* originated from *O. rufipogon* in the Yangtze River region of China (Oka, 1974; Londo et al., 2006; Huang et al., 2012; Sun et al., 2015), and the divergence between *indica* and *japonica* subspecies occurred 0.4 Mya (Kumagai et al., 2010; Chen et al., 2012). Our phylogenetic analysis showed that *O. nivara* and *O. rufipogon* were present in two separate branches, supporting the evolutionary model of multiple origins. LM8 was originated approximately 0.32 Mya and harbors morphological characteristics specific to wild rice such as shattering, hard glumes, and small grains (**Figure 5**). Thus, it could be concluded that LM8 is a kind of *japonica*-type weedy rice from a cross between *japonica* and wild rice, which confirmed the result of taxonomic study.

Chromosome-level genome assemblies may generally accelerate gene discovery in crops to improve yield, quality, and disease resistance (Rao et al., 2014; Qian et al., 2016; Bai et al., 2018). As genome assemblies of Asian cultivated rice varieties such as MH63, ZH97, and R498 become available, a large number of structural variations have been successfully obtained, which would have a wide-range impact on crop genetic improvement (Zhang et al., 2016; Du et al., 2017). For example, Zhang et al. (2014) assembled five AA-genome rice species and identified 14 PSGs that are closely related to rice flowering, development, reproduction, biotic and abiotic resistance through comparative genomics analyses. Although many genomes have been assembled in the *Oryza* genus, only one of them belongs to weedy rice (WRAH), which was used to discuss the origin of weedy rice (Sun et al., 2019). In this study, we reported another weedy rice (LM8) genome for the purpose of identifying genes. This chromosome-level genome assembly contains 672 unique genes specific to weedy rice compared with other 17 *Oryza* genomes (**Figure 5**). Besides, the comparison of the contig N50 (6.09 Mb in WRAH and 17.86 Mb in LM8) and sequence gaps (94 in WRAH and 25 in LM8; **Table 1**) between these two weedy rice genomes (Sun et al., 2019) indicates the high-quality LM8 genome assembly is able to serve as a reference for accelerating the identification of genes from weedy rice, thus improving the cultivated rice varieties.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available. This data can be found here: National Center for Biotechnology Information (NCBI) BioProject database under accession number PRJNA754271.

AUTHOR CONTRIBUTIONS

FL performed the experiments. FL and ZH conducted data analyses and wrote the manuscript. WQ contributed to construct of genetic population and experimental guidance. JW, YS, YCu, JL, JG, DLo, and WF contributed to help data analyses. DLi, BN, ZZ, YCh, and LZ contributed to the material preparation, collection, and measurement. QY and XZ designed the experiment and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (31970237), the Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences (2060302-2), the Sanya Yazhou Bay Science and Technology

City (SKJC-2020-02-001), and the Fundamental Research Funds (S2021ZD01).

ACKNOWLEDGMENTS

We acknowledge the Anhui Academy of Agricultural Sciences for kindly providing the cultivated rice variety "Shen 08S." We appreciate the linguistic assistance provided by TopEdit (www.topeditsci.com) during preparation of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.775051/full#supplementary-material

REFERENCES

Agarwal, G., Clevenger, J., Pandey, M. K., Wang, H., Shasidhar, Y., Chu, Y., et al. (2018). High-density genetic map using whole-genome re-sequencing for fine mapping and candidate gene discovery for disease resistance in peanut. *Plant Biotechnol. J.* 16, 1954–1967. doi: 10.1111/pbi.12930

Bai, S., Yu, H., Wang, B., and Li, J. (2018). Retrospective and perspective of rice breeding in China. *J. Genet. Geno.* 45, 603–612. doi: 10.1016/j.jgg.2018.10.002

Baker, H. G. (1974). The evolution of weeds. *Annu. Rev. Ecol. Syst.* 5, 1–24. doi: 10.1146/annurev.es.05.110174.000245

Bao, W., Kojima, K. K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 6:11. doi: 10.1186/s13100-015-0041-9

Basunia, M. A., Nonhebel, H. M., Backhouse, D., and Mcmillan, M. (2021). Localised expression of OsIAA29 suggests a key role for auxin in regulating development of the dorsal aleurone of early rice grains. *BioRxiv [prperint]*. doi: 10.1007/s00425-021-03688-z

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Cali, D. S., Kim, J. S., Ghose, S., Alkan, C., and Mutlu, O. (2018). Nanopore sequencing technology and tools: computational analysis of the current state, bottlenecks, and future directions. *Brief. Bioinform.* 20, 1542–1559. doi: 10.1093/bib/bby017

Chang, T. T. (1976). The origin, evolution, cultivation, dissemination, and diversification of Asian and African rices. *Euphytica* 25, 425–441. doi: 10.1007/BF00041576

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., et al. (2012). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza genome* evolution. *Nat. Commun.* 4:1595. doi: 10.1038/ncomms2596

Chen, S., Zhou, Y., Zhou, Chen, Y., and Gu, J. (2018). Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560

Cho, Y., Chung, T., and Suh, H. (1995). Genetic characteristics of Korean weedy rice (*Oryza sativa* L.) by RFLP analysis. *Euphytica* 86, 103–110. doi: 10.1007/BF00022015

Cui, Y., Song, B., Li, L., Li, Y., Huang, Z., Caicedo, A. L., et al. (2016). Little white lies: pericarp color provides insights into the origins and evolution of Southeast Asian weedy rice. *G3: Genes Genomes Genet* 6, 4105–4114. doi: 10.1534/g3.116.035881

Dai, L., Dai, W., Song, X., Lu, B., and Qiang, S. (2013). A comparative study of competitiveness between different genotypes of weedy rice (*Oryza sativa*) and cultivated rice. *Pest Manage. Sci.* 70, 113–122. doi: 10.1002/ps.3534

De Bie, T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 22, 1269–1271. doi: 10.1093/bioinformatics/btl097

Du, H., Yu, Y., Ma, Y., Gao, Q., Cao, Y., Chen, Z., et al. (2017). Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* 8, 15324–15336. doi: 10.1038/ncomms15324

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., et al. (2006). GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical Appl. Genet.* 112, 1164–1171. doi: 10.1007/s00122-006-0218-1

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* 296, 92–100. doi: 10.1126/science.1068275

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33, D121–D124. doi: 10.1093/nar/gki081

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, J. R. K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Han, B., Xue, Y., Li, J., Deng, X., and Zhang, Q. (2007). Rice functional genomics research in China. *Philos. Trans. R. Soc. Lond. Ser. B: Biol. Sci.* 362, 1009–1021. doi: 10.1098/rstb.2007.2030

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36, 2253–2255. doi: 10.1093/bioinformatics/btz891

Hu, Z., Deng, G., Mou, H., Xu, Y., Chen, L., Yang, J., et al. (2018). A re-sequencing-based ultra-dense genetic map reveals a gummy stem blight resistance-associated gene in Cucumis melo. *DNA Res.* 225, 1–10. doi: 10.1093/dnares/dsx033

Huang, P., Molina, J., Flowers, J. M., Rubinstein, S., Schaal, B. A., Purugganan, M. D., et al. (2012). Phylogeography of Asian wild rice, *Oryza rufipogon*: a genome-wide view. *Mol. Ecol.* 21, 4593–4604. doi: 10.1111/j.1365-294X.2012.05625.x

Huang, X., Qian, Q., Liu, Z., Sun, H., He, S., Luo, D., et al. (2009). Natural variation at the DEP1 locus enhances grain yield in rice. *Nat. Genet.* 41, 494–497. doi: 10.1038/ng.352

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695

Huang, X., Yang, S., Gong, J., Zhao, Y., Feng, Q., Gong, H., et al. (2015). Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat. Commun.* 6, 6258–6267. doi: 10.1038/ncomms7258

Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., et al. (2011). Genome-wide association study of flowering time and grain yield traits in a worldwide collectionof rice germplasm. *Nat. Genet.* 44, 32–39. doi: 10.1038/ng.1018

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215. doi: 10.1093/nar/gkn785

International Wheat Genome Sequencing Consortium [IWGSC] (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345:1251788. doi: 10.1126/science.1251788

Ishikawa, R., Toki, N., Imai, K., Sato, Y. I., Yamagishi, H., Shimamoto, Y., et al. (2005). Origin of weedy rice grown in bhutan and the force of genetic diversity. *Genet. Resour. Crop Evol.* 52, 395–403. doi: 10.1007/s10722-005-2257-x

Jacquemin, J., Ammiraju, J. S. S., Haberer, G., Billheimer, D. D., Yu, Y., Liu, L. C., et al. (2014). Fifteen million years of evolution in the *Oryza genus* shows extensive gene family expansion. *Mol. Plant* 7, 642–656. doi: 10.1093/mp/sst149

Jens, K., Michael, W., Jessica, E., Martin, H. S., Jan, G., and Frank, H. (2016). Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res.* 44:e89. doi: 10.1093/nar/gkw092

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khush, G. S. (2001). Green revolution: the way forward. *Nat. Rev. Genet.* 2, 815–822. doi: 10.1038/35093585

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kumagai, M., Wang, L., and Ueda, S. (2010). Genetic diversity and evolutionary relationships in genus Oryza revealed by using highly variable regions of chloroplast DNA. *Gene* 462, 44–51. doi: 10.1016/j.gene.2010.04.013

Langfelder, P., and Horvath, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* 46:i11. doi: 10.18637/jss.v046.i11

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

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

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

Li, L., Stoeckert, C. J. Jr., and Roots, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Li, X., Singh, J., Qin, M., Li, S., Zhang, X., Zhang, M., et al. (2019). Development of an integrated 200K SNP genotyping array and application for genetic mapping, genome assembly improvement and genome wide association studies in pear (Pyrus). *Plant Biotechnol. J.* 17, 1582–1594. doi: 10.1111/pbi.13085

Li, X., Wu, L., Wang, J., Sun, J., Xia, X., Geng, X., et al. (2018). Genome sequencing of rice subspecies and genetic analysis of recombinant lines reveals regional yield- and quality-associated loci. *BMC Biol.* 16:102. doi: 10.1186/s12915-018-0572-x

Li, Y., Fan, C., Xing, Y., Jiang, Y., Luo, L., Sun, L., et al. (2011). Natural variation in GS5 plays an important role in regulating grain size and yield in rice. *Nat. Genet.* 43, 1266–1269. doi: 10.1038/ng.977

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* 35, 62–67. doi: 10.1016/S0925-4005(96)02015-1

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444

Londo, J. P., Chiang, Y., Hung, K., Chiang, T. Y., and Schaal, B. A. (2006). Phylogeography of Asian wild rice, *Oryza rufipogon*, reveals multiple independent domestications of cultivated rice, *Oryza sativa*. *Proc. Natl. Acad. Sci. U S A.* 103, 9578–9583. doi: 10.1073/pnas.0603152103

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Lu, B., Narede, M. E. B., Juliano, A. B., and Jackson, M. T. (2000). *Preliminary Studies on Taxonomy and Biosystematics of the AA Genome Oryza species (Poaceae)*. Clayton, CSU: CSIRO Publishing.

Luo, J., Liu, H., Zhou, T., Gu, B., Huang, X., Shangguan, Y., et al. (2013). An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. *Plant Cell* 25, 3360–3376. doi: 10.1105/tpc.113.113589

Luo, X., Xu, L., Wang, Y., Dong, J., Chen, Y., Tang, M., et al. (2020). An ultra-high-density genetic map provides insights into genome synteny, recombination landscape and taproot skin colour in radish (*Raphanus sativus* L.). *Plant Biotechnol. J.* 18, 274–286. doi: 10.1111/pbi.13195

Ma, X., Fu, Y., Zhao, X., Jiang, L., Zhu, Z., Gu, P., et al. (2016). Genomic structure analysis of a set of *Oryza nivara* introgression lines and identification of yield-associated QTLs using whole-genome resequencing. *Sci. Rep.* 6, 27425–27437. doi: 10.1038/srep27425

Mao, H., Sun, S., Yao, J., Wang, C., Yu, S., Xu, C., et al. (2018). Linking differential domain functions of the GS3 protein to natural variation of grain size in rice. *PNAS.* 107, 19579–19584. doi: 10.1073/pnas.1014419107

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

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

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., et al. (2016). A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* 13, 587–590. doi: 10.1038/nmeth.3865

Nadir, S., Khan, S., Zhu, Q., Henry, D., Wei, L., Lee, D. S., et al. (2018). An overview on reproductive isolation in *Oryza sativa* complex. *AOB Plants* 10, 60–73. doi: 10.1093/aobpla/ply060

Nikolau, B. J., Ohlrogge, J. B., and Wurtele, E. S. (2003). Plant biotin-containing carboxylases. *Arch. Biochem. Biophys.* 414, 211–222. doi: 10.1016/S0003-9861(03)00156-5

Oka, H. I. (1974). Experimental studies on the origin of cultivated rice. *Genetics* 78, 475–486. doi: 10.1093/genetics/78.1.475

Ooijen, J. V., Ooijen, J. W. V., Ooijen, J., Hoorn, J., Duin, J., and Verlaat, J. V. T. (2009). *MapQTL® 6. Software for the Mapping of Quantitative Trait Loci in Experimental Populations of Diploid Species*. Wageningen: Kyazma BV.

Peng, X., Liu, H., Chen, P., Tang, F., Hu, Y., Wang, F., et al. (2019). A chromosome-scale genome assembly of paper mulberry (*Broussonetia papyrifera*) provides new insights into its forage and papermaking usage. *Mol. Plant* 12, 661–677. doi: 10.1016/j.molp.2019.01.021

Puttick, M. N. (2019). MCMCtreeR: functions to prepare MCMCtree analyses and visualize posterior ages on trees. *Bioinformatics* 35, 5321–5322. doi: 10.1093/bioinformatics/btz554

Qi, P., Lin, Y. S., Song, X. J., Shen, J. B., Huang, W., Shan, J. X., et al. (2012). The novel quantitative trait locus GL3.1 controls rice grain size and yield by regulating Cyclin-T1;3. *Cell Res.* 22, 1666–1680. doi: 10.1038/cr.2012.151

Qian, Q., Guo, L., Smith, S. M., and Li, J. (2016). Breeding high-yield superior quality hybrid super rice by rational design. *Natl. Sci. Rev.* 3, 283–294. doi: 10.1093/nsr/nww006

Rao, Y., Li, Y., and Qian, Q. (2014). Recent progress on molecular breeding of rice in China. *Plant Cell Rep.* 33, 551–564. doi: 10.1007/s00299-013-1551-x

Reisner, W., Larsen, R. B., Silahtaroglu, R., Kristensen, R., Tommerup, N., Tegenfeldt, R. O., et al. (2010). Single-molecule denaturation mapping of DNA in nanofluidic channels. *PNAS.* 107, 13294–13299. doi: 10.2307/25708710

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

Sharma, B., and Taganna, J. (2020). Genome-wide analysis of the U-box E3 ubiquitin ligase enzyme gene family in tomato. *Sci. Rep.* 10:9581. doi: 10.1038/s41598-020-66553-1

Shinobu, N., Junko, S., Rieko, O., Kana, H., Rika, Y., Noriyuki, K., et al. (2002). Search for and analysis of single nucleotide polymorphisms (SNPs) in rice (*Oryza sativa*, *Oryza rufipogon*) and establishment of SNP markers. *DNA Res.* 9, 163–171. doi: 10.1093/dnares/9.5.163

Shivrain, V. K., Burgos, N. R., Sales, M. A., and Yong, Y. I. (2010). Polymorphisms in the ALS gene of weedy rice (*Oryza sativa* L.) accessions with differential tolerance to imazethapyr. *Crop Protect.* 29, 336–341. doi: 10.1016/j.cropro.2009.10.002

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., et al. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* 40, 1023–1028. doi: 10.1038/ng.169

Simão, F. A., Waterhouse, R. M., Panagiotis, I., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Song, X., Huang, W., Shi, M., Zhu, M. Z., and Lin, H. X. (2007). A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase. *Nat. Genet.* 39, 623–630. doi: 10.1038/ng2014

Spannagl, M., Nussbaumer, T., Bader, K. C., Martis, M. M., Seidel, M., Kugler, K. G., et al. (2016). PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* 44, D1141–D1147. doi: 10.1093/nar/gkv1130

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: join map. *Plant J.* 3, 739–744. doi: 10.1111/j.1365-313X.1993.00739.x

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637–644. doi: 10.1093/bioinformatics/btn013

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19, 215–225. doi: 10.1093/bioinformatics/btg1080

Stein, J. C., Yu, Y., Copetti, D., Zwickl, D. J., Zhang, L., Zhang, C., et al. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* 7, 285–296. doi: 10.1038/s41588-018-0040-0

Sun, C., Wang, X., Yoshimura, A., and Doi, K. (2002). Genetic differentiation for nuclear, mitochondrial and chloroplast genomes in common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 104, 1335–1345. doi: 10.1007/s00122-002-0878-4

Sun, J., Ma, D., Tang, L., Zhao, M., Zhang, G., Wang, W., et al. (2019). Population genomic analysis and de novo assembly reveal the origin of weedy rice as an evolutionary game. *Mol. Plant* 12, 632–647. doi: 10.1016/j.molp.2019.01.019

Sun, J., Qian, Q., Ma, D. R., Xu, Z. J., Liu, D., Du, H. B., et al. (2013). Introgression and selection shaping the genome and adaptive loci of weedy rice in northern China. *New Phytol.* 197, 290–299. doi: 10.1111/nph.12012

Sun, X., Jia, Q., Guo, Y., Zheng, X., and Liang, K. (2015). Whole-genome analysis revealed the positively selected genes during the differentiation of indica and temperate japonica rice. *PLoS One* 10:e0119239. doi: 10.1371/journal.pone.0119239

Vaughan, D. A., Balázs, E., and Heslop-Harrison, J. S. (2007). From crop domestication to super-domestication. *Ann. Bot.* 100, 893–901. doi: 10.1093/aob/mcm224

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, C., Yu, H., Huang, J., Wang, W., Faruquee, M., Zhang, F., et al. (2020). Towards a deeper haplotype mining of complex traits in rice with RFGB v2.0. *Plant Biotechnol J.* 18, 14–16. doi: 10.1111/pbi.13215

Wang, F., Zhang, J., Chen, Y., Zhang, C., Gong, J., Song, Z., et al. (2020). Identification of candidate genes for key fibre-related QTLs and derivation of favourable alleles in *Gossypium hirsutum* recombinant inbred lines with *G.barbadense* introgressions. *Plant Biotechnol. J.* 18, 707–720. doi: 10.1111/pbi.13237

Wang, S., Wu, K., Yuan, Q., Liu, X., Liu, Z., Lin, X., et al. (2012b). Control of grain size, shape and quality by OsSPL16 in rice. *Nat. Genet.* 44, 950–954. doi: 10.1038/ng.2327

Wang, S., Basten, C. J., and Zeng, Z. (2012a). *Windows QTL Cartographer v2.5. Department of Statistics*. Raleigh, NC: North Carolina State University.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012c). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40:e49. doi: 10.1093/nar/gkr1293

Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., et al. (2018). Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557, 43–49. doi: 10.1038/s41586-018-0063-9

Wei, C., Gao, Y., Xie, W., Liang, G., Kai, L., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46, 714–721. doi: 10.1038/ng.3007

Wet, J. M. J. D., and Harlan, J. R. (1975). Weeds and domesticates: evolution in the man-made habitat. *Econ. Bot.* 29, 99–107. doi: 10.1007/BF02863309

Xiao, M., Phong, A., Ha, C., Chan, T. F., Cai, D., Leung, L., et al. (2007). Rapid DNA mapping by fluorescent single molecule detection. *Nuclc Acids Res.* 35:e16. doi: 10.1093/nar/gkl1044

Xu, J., Xing, Y., Xu, Y., and Wan, J. (2021). Breeding by design for future rice: genes and genome technologies. *Crop J.* 9, 491–496. doi: 10.1016/j.cj.2021.03.006

Xun, X., Xin, L., Song, G., Jensen, J. D., Hu, F., Li, X., et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111. doi: 10.1038/nbt.2050

Yang, J., Zhang, C., Zhao, N., Zhang, L., Hu, Z., Chen, S., et al. (2018). Chinese root-type mustard provides phylogenomic insights into the evolution of the multi-use diversified allopolyploid *Brassica juncea*. *Mol. Plant* 11, 512–514. doi: 10.1055/s-0043-120348

Yang, W., Wu, K., Wang, B., Liu, H., Guo, S., Guo, X., et al. (2021). The RING E3 ligase CLG1 targets GS3 for degradation via the endosome pathway to determine grain size in rice. *Mol. Plant* 14, 1699–1713. doi: 10.1016/j.molp.2021.06.027

Yang, X., and Hwa, C. (2008). Genetic modification of plant architecture and variety improvement in rice. *Heredity* 101, 396–404. doi: 10.1038/hdy.2008.90

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., et al. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat. Genet.* 48, 927–934. doi: 10.1038/ng.3596

Yu, J., Hu, S., Wang, J., Li, S., Wong, K. S. G., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science* 296, 79–92. doi: 10.1126/science.1068037

Zdobnov, E. M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848. doi: 10.1093/bioinformatics/17.9.847

Zeng, L., Tu, X.-L., Dai, H., Han, F., Lu, B., Wang, M., et al. (2019). Whole genomes and transcriptomes reveal adaptation and domestication of pistachio. *Genome Biol.* 20:79. doi: 10.1186/s13059-019-1686-3

Zhang, J., Chen, L., Xing, F., Kudrna, D. A., Yao, W., Copetti, D., et al. (2016). Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. U S A.* 113, E5163–E5171. doi: 10.1073/pnas.1611012113

Zhang, Q., Liang, Z., Cui, X., Ji, C., Li, Y., Zhang, P., et al. (2018). N6-Methyladenine DNA methylation in Japonica and indica rice genomes and its association with gene expression, plant development, and stress responses. *Mol. Plant* 11, 1492–1508. doi: 10.1016/j.molp.2018.11.005

Zhang, Q., Zhu, T., Xia, E., Shi, C., Liu, Y., Zhang, Y., et al. (2014). Rapid diversification of five Oryza AA genomes associated with rice adaptation. *Proc. Natl. Acad. Sci. U S A.* 111, 4954–4962. doi: 10.1073/pnas.1418307111

Zhu, Q., and Ge, S. (2005). Phylogenetic relationships among a-genome species of the genus Oryza revealed by intron sequences of four nuclear genes. *New Phytol.* 167, 249–265. doi: 10.1111/j.1469-8137.2005.01406.x

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.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 26 November 2021
doi: 10.3389/fpls.2021.748531

A Rice Ancestral Genetic Resource Conferring Ideal Plant Shapes for Vegetative Growth and Weed Suppression

Noritoshi Inagaki[1†], Hidenori Asami[2,3†], Hideyuki Hirabayashi[4], Akira Uchino[5], Toshiyuki Imaizumi[3] and Ken Ishimaru[4]*

[1] Research Center for Advanced Analysis, National Agriculture and Food Research Organization (NARO), Tsukuba, Japan, [2] Western Region Agricultural Research Center (Kinki, Chugoku, and Shikoku Regions), National Agriculture and Food Research Organization (NARO), Fukuyama, Japan, [3] Institute for Plant Protection, National Agriculture and Food Research Organization (NARO), Tsukuba, Japan, [4] Institute of Crop Science, National Agriculture and Food Research Organization (NARO), Tsukuba, Japan, [5] Central Region Agricultural Research Center (Kanto, Tokai, and Hokuriku Regions), National Agriculture and Food Research Organization (NARO), Tsu, Japan

OPEN ACCESS

Edited by:
Monica Fernandez-Aparicio, Institute for Sustainable Agriculture, Spanish National Research Council (CSIC), Spain

Reviewed by:
Stanley Omar Samonte, Texas A&M AgriLife Research and Extension Center at Beaumont, United States
Yoshihiro Hirooka, Kindai University, Japan

***Correspondence:**
Noritoshi Inagaki
ninagaki@affrc.go.jp

† These authors have contributed equally to this work

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

Received: *28 July 2021*
Accepted: *28 October 2021*
Published: *26 November 2021*

Citation:
Inagaki N, Asami H, Hirabayashi H, Uchino A, Imaizumi T and Ishimaru K (2021) A Rice Ancestral Genetic Resource Conferring Ideal Plant Shapes for Vegetative Growth and Weed Suppression. Front. Plant Sci. 12:748531. doi: 10.3389/fpls.2021.748531

To maximize crop growth, crops need to capture sunlight efficiently. This property is primarily influenced by the shape of the crops such as the angle, area, and arrangement of leaves. We constructed a rice (*Oryza sativa* L.) inbred line that displayed an ideal transition of plant shapes in terms of sunlight receiving efficiency. During vegetative growth, this line exhibited tiller spreading with increased tiller number, which formed a parabolic antenna-like structure. The architecture probably improved light reception efficiency of individuals compared with the recurrent parent. The line achieved not only acceleration of the vegetative growth, but also significant suppression of weed growth under the canopy. The increased light reception efficiency of the line has consequently reduced the amount of incident light to the ground and supplied significant competitiveness against weeds. The spread tillers became erect from the entry of the reproductive growth phase, adaptively sustaining light reception efficiency in thicker stands. The line carries a small chromosomal segment from *Oryza rufipogon* Griff., a putative progenitor of Asian cultivated rice. The introduced chromosome segment had little effect on grain yield and quality. Our results shed light on potentials hidden in the wild rice chromosome segment to achieve the valuable traits.

Keywords: ancestral genetic resource, Asian cultivated rice (*Oryza sativa* L), light reception efficiency, plant shape, tiller angle, vegetative growth, weed control, wild rice (*Oryza rufipogon* Griff.)

INTRODUCTION

Asian cultivated rice, *Oryza sativa* (*O. sativa*), is one of the world's most important crops, sustaining several billion people as a staple food (Global Rice Science Partnership, 2013). To ensure permanence of humanity, improvement of rice cultivation systems should be required not only from economic perspectives, but also from viewpoints based on sustainability.

O. sativa was domesticated from a wild rice species, *Oryza rufipogon* (*O. rufipogon*), which is distributed throughout Asia and Oceania (Chen et al., 2019). Human ancestors empirically selected

beneficial lines along their cropping systems from the progenitor. Such domestication processes substantially narrowed the genetic diversity at genetic bottlenecks formed during these winnowing steps (Doebley et al., 2006). Comprehensive genome sequencing and analyses of *O. rufipogon* and *O. sativa* quantitatively ascertained the vast reduction in genetic diversity that occurred during domestication (Zhu et al., 2007; Huang et al., 2012; Chen et al., 2019). On the other hand, modern breeding has certainly attained some success in constructing elite cultivars using the limited amount of genetic variation still retained in cultivated rice plants and acquired mutations in key genes involved in critical agricultural traits; however, this strategy has been sometimes confronted with limits attributed to a low level of genetic variation. Therefore, the positive resurrection of genetic variations lost during domestication will be a powerful tool to solve deadlocks in genetic improvement (Kamboj et al., 2020).

During the rice domestication, one of the most drastic events was exclusion of prostrate plants, mostly exhibited by wild rice, which triggered to form majority of the erect plants found in almost all the modern rice cultivars. This selection probably preceded at the early phase of the domestication, which conferred an increase in yield per area by supplying the ability for dense planting. This characteristic was extremely important in the ancient times when farmland expanding ability was low and/or farmland with appropriate conditions was restricted.

In 2008, the domestication locus responsible for the erect tiller trait was identified and named the *Prostrate growth 1* (*Prog1*), a gene located on the short arm of chromosome 7 (Jin et al., 2008; Tan et al., 2008). The *Prog1* gene encodes a C_2H_2-type zinc finger transcription factor (Agarwal et al., 2007). Recently, Wu et al. (2018) reported that *O. rufipogon* (accession DXCWR) possesses approximately 110 kbp of an additional chromosomal segment named the *RICE PLANT ARCHITECTURE DOMESTICATION* (*RPAD*) in the near vicinity of the *Prog1* gene. The *RPAD* segment contains seven tandem *Prog1*-like genes; at least three of the *Prog1*-like genes were confirmed to modify plant architecture. The erect habit of cultivated rice plants was presumed to be guided by the synergistic effects of sequence changes in the *Prog1* and deletion of the *RPAD* segment from the *O. rufipogon* genome (Huang et al., 2020).

Growth of crops, including rice plants, certainly requires the adequate sunlight as a basal energy source for photosynthesis and is greatly affected by the efficiency of sunlight reception. In fact, Monteith (1977) experimentally found that primary crop production is proportional to the amount of intercepted light energy. Therefore, improving the light-intercepting characteristic of plant shape is expected to promote crop growth. The exact plant shape, which is most effective in perceiving solar radiation, is a point of debate. Theoretical studies examining the relationship between leaf angles and dry matter production have suggested that erect leaves are preferable for increasing the capture of sunlight and enhancing photosynthetic production during the late vegetative growth phase when leaves grow thicker and become mutually shaded (Monteith, 1965; Duncan et al., 1967). In terms of growth promotion during the late vegetative phase, rice lines possessing erect leaves and/or panicles have been constructed that can achieve some increase in yield (Sakamoto et al., 2006; San et al., 2018; Fei et al., 2019). In contrast, Monteith (1965) and Duncan et al. (1967) also reported that inclined leaves are more efficient in perceiving light when there is no mutual shading as typically found in plants with a low leaf area index (LAI), implying that a spreading phenotype is preferable for plants during the early vegetative growth phase. These interpretations indicate that the ideal plant shape for efficient light capture changes depending on the growth phase of the crop.

Among biotic factors interfering with crop yield, weeds constantly invade and spread in fields and are a major cause of reduced yield in all the regions of the globe (Chauhan, 2020). Conventional weed control by manually removing weeds from the field is burdensome and the frequent application of herbicides is expensive and comes with ecological costs such as environmental destruction risk and/or the appearance of herbicide-resistant weeds. Fewer herbicide applications relieve not only the economic burden, but also the ecological issues; however, these applications alone will reduce yields of crops due to insufficient weed control. More ecological approaches, which can substantially reduce herbicide applications, i.e., construction of cultivars with an appropriate plant shape that preserves strong competitiveness against weeds, have been considered (Johnson et al., 1998; Fischer et al., 2001; Dass et al., 2017).

The ideal plant shape for rice to improve weed competitiveness is a plant with many tillers and sloping leaves (Gibson et al., 2003; Koarai and Morita, 2003). These traits would form a deep canopy, limit light penetration onto the ground, and then suppress the growth of weeds under the canopy. From these perspectives, researchers have sought and examined cultivars displaying such ideal plant shapes for strong weed competitiveness. However, it was feared that such traits would lead to a severe decrease in yield due to energy loss by mutual shading of the crop leaves, i.e., the features conflict with the preferable plant shape for high yields in the late vegetative phase. Therefore, cultivars with compatible plant shape are earnestly demanded, especially in developing countries with little economic capacity, which would reduce the weeding costs. Even in developed countries, the decrease of herbicide application has great merits to relieve not only environmental loads, but also appearance of herbicide-resistant weeds (Annett et al., 2014; Heap and Duke, 2018; Islam et al., 2018).

From the perspective of agricultural management, sparse planting of rice seedlings is demanded to save labor and the production cost. However, the sparse planting brings to a new weakness in weed control because the canopy closure of rice plants, which represses weed growth, is delayed. We have to examine ideal plant shapes for the sparse planting that maintain yield and are compatible with weed control.

In this study, we established the rice near-isogenic line (NIL) of Koshihikari holding ideal plant shapes that are compatible with yield and weed control even under sparse planting condition. Koshihikari is the most widely accepted *japonica* cultivar for staple food in Japan due to its good taste and is recognized to be a monumental existence that has made great contributions to rice breeding in Japan as a mating parent. Koshihikari has erect tillers, which characteristic is generally found in common cultivated

rice. The Koshihikari-background NIL contains a genomic segment of chromosome 7 from an accession of *O. rufipogon* collected in Thailand, which segment carries the novel *Prog1* sequence with the *RPAD*. The NIL displayed the spreading tiller phenotype in the vegetative growth phase; however, the tillers began to erect from the entry of the reproductive growth phase. This transition between the plant shapes complies with maintenance of optimal light receiving efficiency throughout plant development; therefore, the NIL balances promoted growth and significant weed suppression. These results suggest that the ancestral genetic resources used in this study have a great ability to upgrade current rice farming to innovative systems fitting with the United Nations Sustainable Development Goals (SDGs)[1].

MATERIALS AND METHODS

Plant Materials and Field Growth Conditions

In this study, Koshihikari, a cultivar of *japonica* rice (*O. sativa*), was used as the control. We used GP9-7, the NIL for Koshihikari, containing a segment of chromosome 7 from an accession of *O. rufipogon* collected in Thailand (IRGC Acc. No. 104814). This line was selected by marker-assisted breeding from lines obtained in the BC_4F_2 generation of KRIL31 backcrossed to Koshihikari. KRIL31 is a line in the chromosome segment substitution lines (CSSLs) containing chromosomal segments from wild relatives in the background of *japonica* cultivars, which has been constructed in our former work (Hirabayashi et al., 2010). This study was conducted at the Tsukuba-Kannondai test fields for NARO (N-36.0°, E-140.1°, 22 m above sea level) and at the Western Region Agricultural Research Center, Fukuyama test fields for NARO (N-34.5°, E-133.4°, 1 m above sea level). Details of the field conditions in our experiments were described in the text or the legends of the figures. A summary of climate conditions in the years of the field experiments indicates in **Supplementary Table 1**.

Cross-Fertilization and Genotyping

After emasculation of pollen on female plants by soaking panicles in hot water at 43°C for 7 min, mating was accomplished by sprinkling the pollen of male plants onto emasculated female stigmata. DNA was extracted from a small piece of a leaf tip using the DNeasy Plant Mini Kit (QIAGEN, Hilden, Germany) following the instruction of the manufacturer. Genotyping was carried out by PCR using the single sequence repeat (SSR) markers (Temnykh et al., 2000; McCouch et al., 2002; International Rice Genome Sequencing Project, 2005) listed in **Supplementary Table 2**. The PCR mixture (10 μl) consisted of 0.5 μl of template DNA, 5 μl of GoTaq Green Master Mix (Promega, Madison, WI, United States), and 0.6 μl of 10 μM primers. Amplification was performed for 35 cycles of 94°C (30 s), 55°C (30 s), and 72°C (30 s). Amplified DNA products were electrophoresed in 3.0% (w/v) NuSieve 3:1 Agarose Gels (Lonza, Basel, Switzerland, United Kingdom).

[1]https://sdgs.un.org

Growth Analyses

At designed intervals, above ground parts of randomly selected seven plants grown in a paddy field (500 m^2; interplant space = 18 cm × 25 cm; 22.2 plants m^{-2}) were harvested. Leaf area was measured with a leaf area meter (AAM-9, Hayashi Denko Corporation Ltd., Tokyo, Japan) and LAI was calculated. All the samples were oven dried for 2 days prior to measuring dry weight. The crop growth rate (CGR) (g m^{-2} day^{-1}) and net assimilation rate (NAR) (g m^{-2} day^{-1}) were calculated using the following equations:

$$CGR = K\,(W2 - W1)/(t2 - t1) \quad (1)$$

$$NAR = (W2 - W1)/(t2 - t1) \{(ln(A2) - ln(A1))/(t2 - t1)\} \quad (2)$$

where, K indicates the plant density (plants m^{-2}), $W1$ and $W2$ represent the above ground dry weight (g) at times $t1$ and $t2$, respectively, and $A1$ and $A2$ represent the leaf area (m^2) at $t1$ and $t2$, respectively.

Measurement of Tiller Inclination Angle and Vegetation Cover Rate

Tiller inclination angles were measured as the angle from the horizontal to a tiller located on the outermost circumference. The vegetation cover rates were calculated from binarized images of plants taken from directly above as the ratio of pixels corresponding to the plant body to the total number of pixels. The binarized images were generated from the a* signals in the CIELAB (Commission internationale de l'éclairage L* a* b*) color space of the original images converted with the ImageJ software version 1.52 k[2] with a color space converter plugin (LPX color; LPIXEL, Tokyo, Japan).

Assays of Weed Growth in Competition With Rice

In this study, *Echinochloa crus-galli* (L.) Beauv. var. *formosensis* Ohwi (*E. crus-galli* var. *formosensis*), a well-known grass weed in Japan, was used. At 10 days after planting (DAP) of rice seedlings on paddy fields (interplant space = 30 cm × 30 cm; 11.1 plants m^{-2}), *E. crus-galli* var. *formosensis* at the first leaf stage was transplanted to the middle of rice rows. The number of tillers of *E. crus-galli* var. *formosensis* was measured every other week. A total of 10 weed plants per plot (1.8 m × 3.3 m = 5.94 m^2) were examined in three repetitions. In addition, all the *E. crus-galli* var. *formosensis* naturally emerging in the three plots of the field were sampled 10 days before rice harvest and the number of emerging weeds, their tiller number, and their dry matter weight were measured.

Measurements of the Relative Photosynthetic Photon Flux Density

Vertical transitions of the RPPFD in the rice stand were measured on a cloudy day at noon using a quantum sensor (LI-190SB, LI-COR, Lincoln, NE, United States). Simultaneous measurements

[2]https://imagej.nih.gov/ij/

were carried out for locations every 10 cm from the ground surface in the rice stand and above the canopy (1 m from the ground surface) using the measurements taken above the stand as 100%. The measurements were recorded three times and the mean value was used in the analysis.

Transitions of the RPPFD in the rice stand during the growth were measured using a line quantum sensor (366813M; Ollie Corporation Ltd., Osaka, Japan). Simultaneous measurements were carried out in the rice stand (ground surface) and outside the stand (1 m from the ground surface) using measurements taken outside the stand as 100%. Measurements were recorded once a week from transplanting to harvest. The measurements were recorded three times and the mean value was used in the analysis. At the same time, the number of rice tillers was measured for three plots that consist of 10 plants.

Grain Yield Measurement and Eating Quality Tests

A total of 12 individual plants randomly selected from 100 plants grown in a field (4 m × 3.3 m = 13.2 m^2; interplant space = 36 cm × 36 cm; 7.7 plants m^{-2}) were harvested as one plot. The panicles in each plot (three plot replicates) were mechanically threshed, the obtained grains were dried, and the weight of the paddy rice was measured to determine grain yield. Samples for the taste test were obtained from plants growing in the same field as those used for grain yield measurements. Taste tests were conducted at the AiHO Rice Cooking Research Institute (Toyokawa, Aichi, Japan). Protein and amylose contents were measured using a Rice Composition Analyzer (Shizuoka Seiki Corporation Ltd., Fukuroi, Japan). Mido Meter (Toyo Rice, Wakayama, Japan) and Rice Taste Analyzer (Satake Corporation, Higashi-Hiroshima, Japan) were used for objective comparisons of taste-related factors. Values for texture of cooked rice were examined by a Tensipresser (Takemoto Electric Incorporation, Tokyo, Japan).

Statistical Analysis

Statistical analyses were conducted using the statistical computing software R, version 3.6.3[3].

RESULTS AND DISCUSSION

Introductory Description of the Near-Isogenic Line (GP9-7)

For the first trial, we examined the agricultural traits of KRILs (Hirabayashi et al., 2010) in field conditions in which seedlings were planted in sparse (interplant space = 36 cm; 7.7 plants

[3]https://www.r-project.org

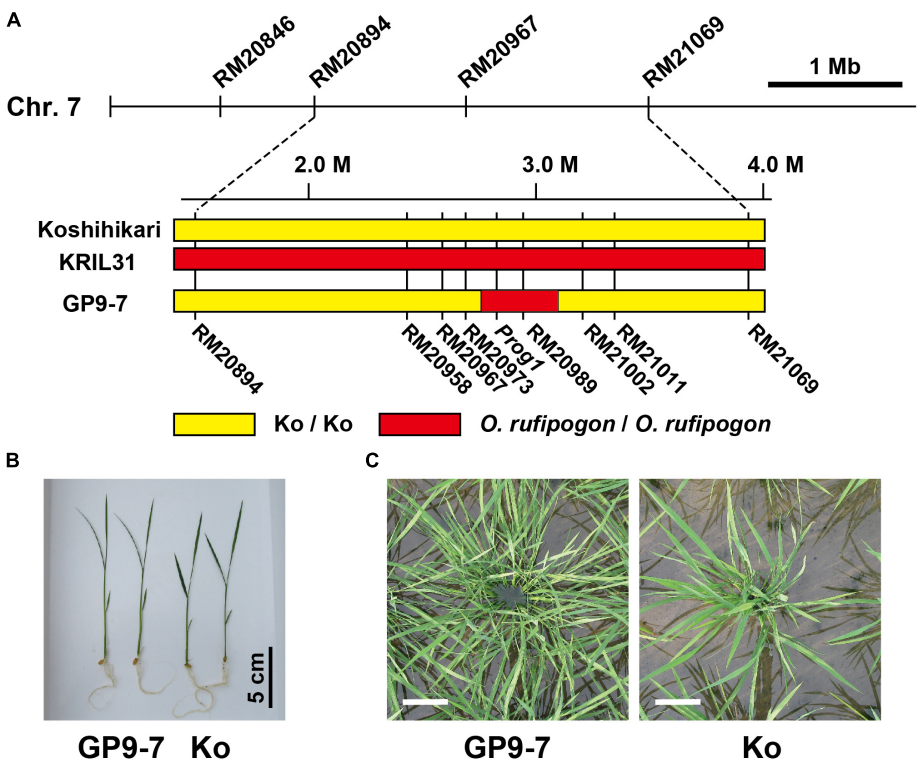


FIGURE 1 | Basic properties of GP9-7. **(A)** A comparative diagram showing the genotyping results of GP9-7, KRIL31, and Koshihikari. The yellow zones indicate the segments derived from Koshihikari and the red zones represent the segments derived from *Oryza rufipogon* (*O. rufipogon*). **(B,C)** Typical plant shapes of the rice plants are used in this study. The photograph **(B)** shows typical nursling seedlings of Koshihikari and GP9-7 at 10 days after sowing. The photographs **(C)** indicate typical plant shapes in paddy fields taken from directly above the plants at 33 days after planting (DAP). Scale bars: 10 cm.

m^{-2}) or dense (interplant space = 18 cm; 30.9 plants m^{-2}) plant densities. The KRILs consist of 40 lines, each of which possess one or few chromosomal segments of *O. rufipogon* (IRGC accession number 104814) in Koshihikari background. This CSSL (KRILs) were constructed to examine whether the wild rice genetic resources can be used to overcome weaknesses and/or add new features to Koshihikari. Among the KRILs, KRIL31 displayed wide spreading with an increased number of tillers, especially in the sparsely planted condition (**Supplementary Figure 1**). Therefore, we focused on these traits that altered vegetative growth presumably caused by a change in their light-receiving posture, which could also provide strong weed competitiveness. Interestingly, KRIL31 did not have a clear dwarf phenotype, even though its tiller number had more than doubled compared with that of Koshihikari. The BC_1F_1 plants derived from a cross between KRIL31 and Koshihikari showed both the traits, indicating these traits were inherited dominantly (data not shown). KRIL31 possesses almost all of the long arm of chromosome 9 and three small segments in chromosomes 1, 3, and 7 from *O. rufipogon* (Hirabayashi et al., 2010; **Supplementary Figure 2A**). KRIL31 tillers maintained a strong spreading trait throughout plant development. When the chromosome 9 segment from *O. rufipogon* replaced that of Koshihikari, the spreading trait in the reproductive phase disappeared (data not shown). Yu et al. (2007) indicated that the *Tiller Angle Control 1* (*TAC1*) gene, which mediates tiller angle during the reproductive stage (late compact stage), is localized in the long arm of chromosome 9. This study aligns with our observation that when this segment was replaced with Koshihikari, resultant progenies lost tiller inclination during the reproductive growth phase. In contrast, the segment of chromosome 7 from *O. rufipogon* was linked to leaning tillers only during the vegetative growth phase. To define the segment responsible for the spreading of tillers, we conducted phenotypic and genotypic analyses of 430 lines of the BC_2F_2 generation, which indicated that the loci for spreading and increasing tiller number were detected between two SSR markers, RM20967 and RM20999 (**Supplementary Figure 2B**). We selected the NIL, GP9-7 from 300 lines of the BC_3F_4 generation by genotypic analyses based on the above information. The chromosomes of GP9-7 had almost reverted to Koshihikari form except for the region between two SSR markers, RM20973 and RM21002 (**Figure 1A**). We used this line in all the subsequent studies.

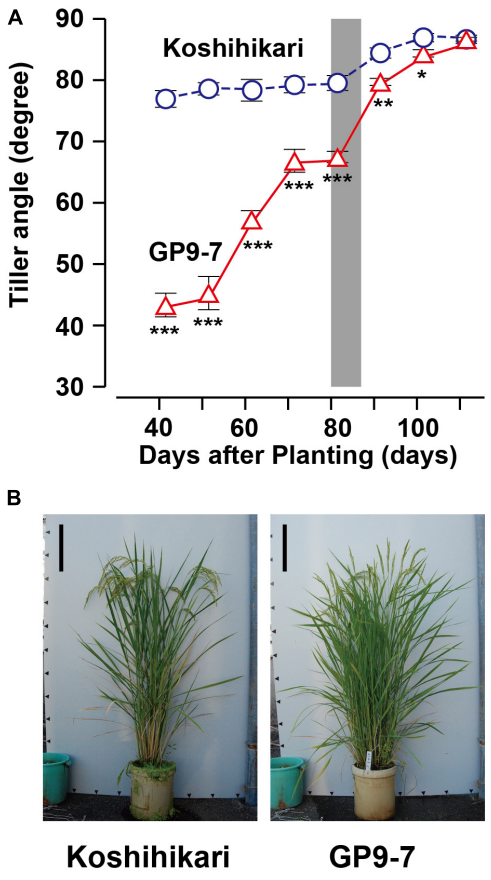


FIGURE 2 | Basic properties of GP9-7 toward the reproductive growth phase. **(A)** Developmental transitions of tiller angles from horizontal in Koshihikari (blue circles with broken lines) and GP9-7 (red triangles with solid lines) grown in a test field at Tsukuba, Japan in the summer of 2020 (planting date: May 21, 2020). The tiller angles of the same six plants were continuously measured and the mean values with SE were shown in the graph. The period for heading is indicated by a vertical gray bar. Symbols, ***, **, and * indicate statistically significant differences compared to Koshihikari at $P < 0.001$, $P < 0.01$, and $P < 0.05$, respectively, calculated by the Welch's *t*-test. The photographs **(B)** display typical plant shapes of Koshihikari and GP9-7 during the ripening phase (25 days after heading). Scale bars: 20 cm.

Inserted Chromosomal Segment of GP9-7

As mentioned above, GP9-7 contains a small segment of chromosome 7 from *O. rufipogon* (IRGC accession number 104814) between two SSR markers, RM20973 and RM21002 (**Figure 1A**). The segment between the two markers corresponds to 358 kbp based on the Nipponbare genome sequence (International Rice Genome Sequencing Project, 2005). This segment includes the *Prog1* gene that is involved in key processes required for domestication and has been reported to determine tiller inclination angle and number of tillers (Jin et al., 2008; Tan et al., 2008). The *Prog1* gene encodes a single C_2H_2 zinc-finger transcription factor (Agarwal et al., 2007; **Supplementary Figure 3**). The *Prog1* sequences of *O. rufipogon* collected in China have been sequenced comprehensively, but the *Prog1* sequence of GP9-7, which is from an accession of *O. rufipogon* (IRGC accession number 104814) collected in Thailand, is novel. This sequence is similar to that of cultivated rice with a deletion of 9 bp in the open reading frame that causes a deletion of three residues near an EAR (ethylene-responsive element binding factor-associated amphiphilic repression)-like motif (Kagale et al., 2010; **Supplementary Figure 3**). Wu et al. (2018) proposed that the *RPAD* region is also involved in determining plant shape, located on the vicinity of the *Prog1* gene in the *O. rufipogon* genome. To verify the insertion of the *RPAD* segment in GP9-7, a PCR assay was performed using a primer

set to detect the *RPAD* insertion. The assay revealed that GP9-7 probably possesses the *RPAD* segment (**Supplementary Figure 4**) that is likely to confer the ability to alter plant shape together with the *Prog1* gene.

Tiller Inclination and Tiller Number in GP9-7

Nursling seedlings of GP9-7 were slightly thinner and more elongated compared with Koshihikari (**Figure 1B**). After planting on paddy fields in sparse condition (interplant space = 36 cm × 36 cm; 7.7 plants m^{-2}), the number of GP9-7 tillers increased and began leaning more each day (**Supplementary Figure 5**; **Supplementary Movie 1**), comparable with previously reported inbred lines containing similar chromosomal segments of *O. rufipogon* (Jin et al., 2008; Tan et al., 2008). The number of GP9-7 tillers was roughly equivalent to those of Koshihikari until about 15 DAP; however, afterward, the number of GP9-7 tillers increased significantly, reaching more than three times those of Koshihikari by 43 DAP (**Supplementary Figure 5**). The plastochron of GP9-7 was almost equivalent to that of Koshihikari (data not shown). This observation and the biased increase in tiller number toward the later developmental stages suggested that higher order tillers that

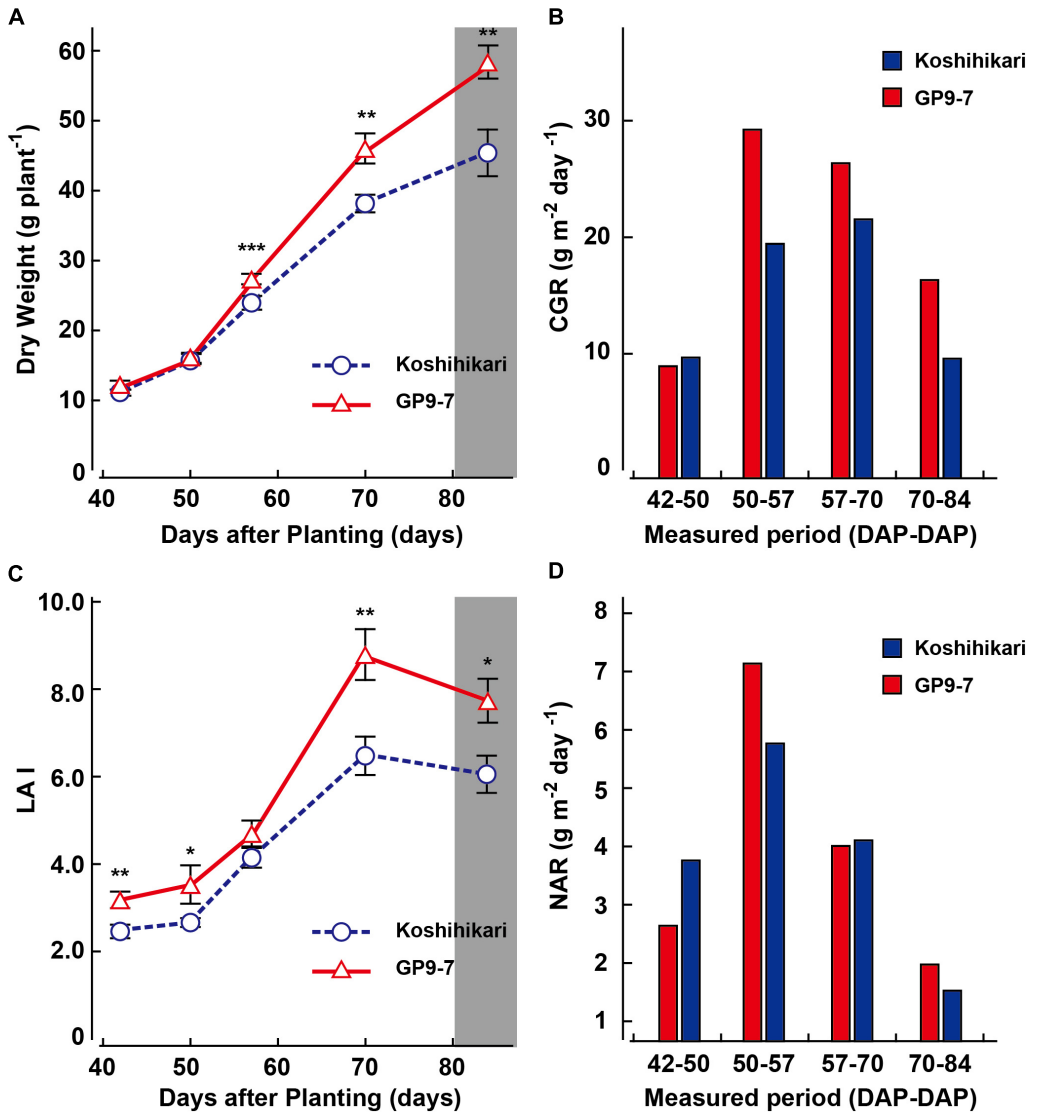


FIGURE 3 | Developmental transitions of growth traits of Koshihikari (blue circles with broken lines or blue bars) and GP9-7 (red triangles with solid lines or red bars) from the vegetative to the reproductive growth phase. Transitions of dry weight per plant **(A)**, crop growth rate (CGR) **(B)**, leaf area index (LAI) **(C)**, and net assimilation rate (NAR) **(D)** of Koshihikari and GP9-7 grown in a test field at Tsukuba, Japan in the summer of 2018 (planting date: May 9, 2018). Mean values of the dry weights and leaf area indexes obtained from seven plants are plotted. The SE values are shown as error bars in the graphs. The period for heading is indicated by a vertical gray bar. Symbols, ***, **, and * indicate statistically significant differences compared to Koshihikari at $P < 0.001$, $P < 0.01$, and $P < 0.05$, respectively, calculated by the Welch's *t*-test.

normally do not appear in Koshihikari did emerge from GP9-7. Tillers that elongate too much can lead to mutual shading, but the emerging GP9-7 tillers spread in all the directions with very few overlaps and then assumed a parabolic antenna-like structure (**Figure 1C**). Therefore, leaves on the tillers of GP9-7 did not compete with each other for light.

The erect tiller trait in cultivated rice, including Koshihikari, is caused by the upward curving of laterally emerging non-elongated internodes (**Supplementary Figure 6A**). Tan et al. (2008) reported that the curved internodes are due to their asymmetric development in which the near-ground border, the outermost cell layer of the tiller base closest to the ground, is longer than the border further away from the ground. Cell sizes for both the borders were almost equivalent suggesting that the curved internodes are due to an increase in cell number on the near-ground border. In contrast, the laterally emerging non-elongated internodes of GP9-7 were straight (**Supplementary Figure 6A**), probably due to symmetric development between the near-ground and the far-ground borders. This observation does not mean that GP9-7 is not gravitropic. Coleoptiles of dark-germinated Koshihikari and GP9-7 seedlings were clearly gravitropic in an experiment in which seedlings were

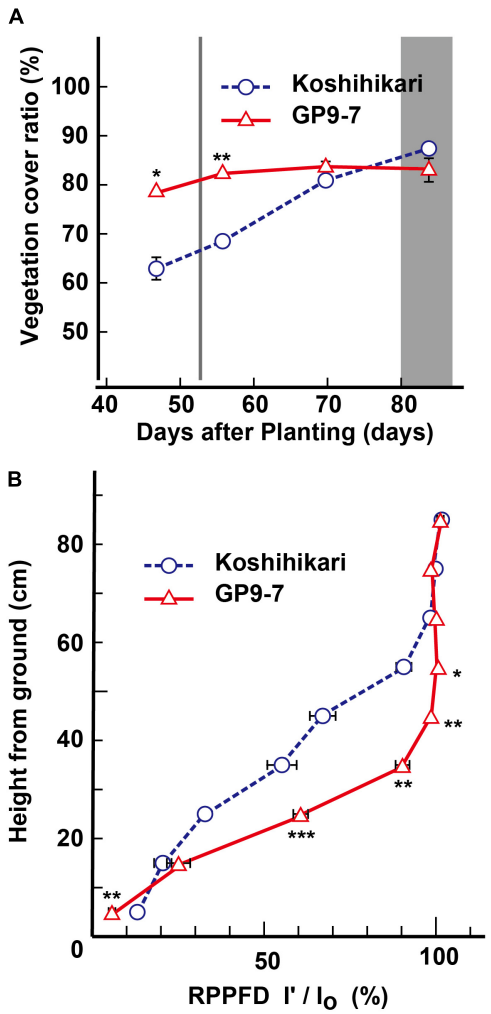


FIGURE 4 | Light reception-related properties of GP9-7. **(A)** Developmental transitions in vegetative coverage of Koshihikari and GP9-7 in the test field at Tsukuba, Japan in the summer of 2018 (planting date: May 28, 2018). The period for heading is indicated by a vertical gray bar. This analysis was performed according to Materials and Methods, using images continuously taken at the same three points in the field, and the mean values with SE were indicated in the graph. **(B)** Vertical transitions of the relative photosynthetic photon flux density (RPPFD) of Koshihikari and GP9-7 measured on the 52nd DAP that is shown by a vertical gray line in panel **(A)**. Symbols, ***, **, and * indicate statistically significant differences compared to Koshihikari at $P < 0.001$, $P < 0.01$, and $P < 0.05$, respectively, calculated by the Welch's *t*-test. Values for Koshihikari are indicated with blue circles with broken lines and values for GP9-7 are represented with red triangles with solid lines.

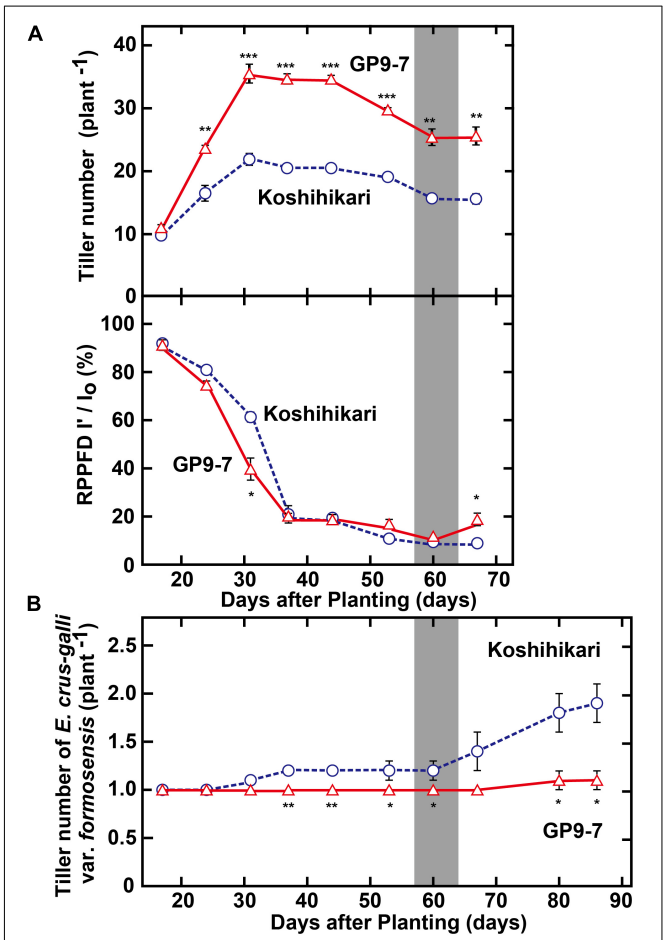


FIGURE 5 | Weed suppression-related properties of GP9-7. **(A)** Developmental transitions in tiller numbers and the RPPFD at ground level of Koshihikari and the GP9-7 in the test field in Fukuyama, Japan in the summer of 2019 (planting date: June 15, 2019). The periods for heading are indicated by vertical gray bars. **(B)** Developmental transitions in tiller numbers of the transplanted weed, *E. crus-galli* var. *formosensis* under the canopy formed by rice stands. Symbols, ***, **, and * indicate statistically significant differences compared to Koshihikari at $P < 0.001$, $P < 0.01$, and $P < 0.05$, respectively, calculated by the Student's *t*-test. Values for Koshihikari are indicated with blue circles with broken lines and values for GP9-7 are represented with red triangles with solid lines.

rotated horizontally at the midpoint of an incubation period (**Supplementary Figure 6B**). Furthermore, tillers of GP9-7 began to rise during the transition from the vegetative to the reproductive growth phase (**Figure 2A**; **Supplementary Movie 2**). This trait was derived from a clear gravitropic bend in GP9-7 nodes, especially the node between internodes III and IV (**Supplementary Figure 6C**), resulting in erect panicles (**Figure 2B**). This trait favorably increasing the light reception efficiency in the thicker stand as is often the case in the reproductive growth phase (Sakamoto et al., 2006; San et al., 2018; Fei et al., 2019). In addition, this trait is preferable for modern agricultural harvesting methods using combine harvesters. As mentioned above, GP9-7 exhibited a clear transition in plant shape from the vegetative growth phase during which plants had inclined tillers to the reproductive growth phase when the tillers were erect.

Accelerated Growth of GP9-7

GP9-7 and Koshihikari were grown in a field in Tsukuba, a region in eastern Japan, using local agricultural practices (interplant space = 18 cm × 25 cm; 30.9 plants m^{-2}). Leaf area and shoot dry matter weight were measured at intervals; LAI, CGR, and NAR were calculated from these values (**Figure 3**). Early vegetative growth of GP9-7 up to 50 DAP was indistinguishable from Koshihikari in terms of dry matter weight per plant; however, beginning at 57 DAP, the increase in dry matter weight of GP9-7 was significantly accelerated (**Figure 3A**). Therefore, the CGR values for GP9-7 during these periods were higher than those of Koshihikari (**Figure 3B**). Since the CGR is recognized as the product of LAI and NAR, it is easy to distinguish which factor contributes more to the increase in the CGR. The factor that promoted the CGR of GP9-7 from 50 to 57 DAP was attributed to a temporary increase in the NAR during the same period (**Figures 3B,D**). In contrast, the higher CGR in the subsequent period (57 to 85 DAP) was due to an increase in LAI (**Figures 3B,C**). The temporary increase in the NAR may be due to improved light-intercepting characteristics in the stand rather than from an increase in the photosynthetic activity of leaves, as discussed in more detail in the next section.

Light Intercepting Ability of GP9-7

The increased growth performance of GP9-7 is probably due to the enhanced amount of light energy received by the individuals attributed to their distinctive shape. Tiller emergence of GP9-7 plants was greater than that of Koshihikari and the tillers were arranged radially causing less mutual shielding (**Figure 1C**; **Supplementary Figure 5**). Therefore, the vegetation cover ratios of GP9-7 in the midvegetative phase were significantly higher than those of Koshihikari (**Figure 4A**). This result suggests that the light-intercepting efficiency of GP9-7 was increased by the plant shape. The RPPFD of GP9-7 in the stand at 53 DAP had a characteristic vertical profile (**Figure 4B**) when the NAR of GP9-7 was at its peak (**Figure 3D**). The vertical RPPFD profile in cultivated rice plants with erect tillers usually indicates a gradual decrease in light intensity from the canopy to the ground surface. In alignment with this expectation, the vertical RPPFD of Koshihikari gradually decreased below 60 cm (**Figure 4B**). In contrast, an extreme decrease in the RPPFD was detected below 40 cm in GP9-7 (**Figure 4B**). This characteristic was attributed to the parabolic antenna-like morphology of GP9-7 that had many uniformly spread tillers with a constant slope, forming wide shadows without mutual shielding (**Figure 1C**).

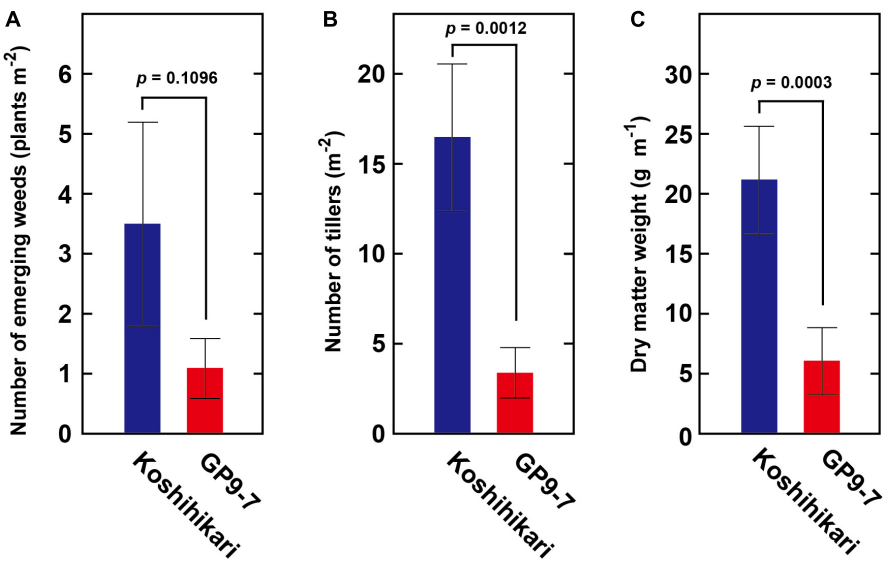


FIGURE 6 | The weed competitiveness of GP9-7 was evaluated by measuring the growth of naturally occurring weeds in the summer of 2019 (planting date: June 15, 2019). The number of emerging weeds **(A)**, their tiller number **(B)**, and the dry matter weight **(C)** of naturally occurring *E. crus-galli* var. *formosensis* under the canopies of Koshihikari or GP9-7 are indicated. Mean values obtained from three plots for Koshihikari and GP9-7 with the SE values are indicated with blue and red bars, respectively. The *p*-values were calculated by the Student's *t*-test and are reported above the bar graphs.

Under the condition, slightly excessive LAI should be connected to effective light capture. In addition, tilted leaf sheaths of GP9-7 were fully exposed to the sky (**Figure 1C**), which implies that photosynthetic assimilation in the leaf sheaths contributes growth promotion of GP9-7. Actually, Guo et al. (2011) reported that rice leaf sheaths had active photosynthetic apparatus such as leaf blades. Moreover, the leaf sheath photosynthesis was accounted for 10 to 20% of the final yield. Our observation that leaf sheaths of GP9-7 were dark green, such as active photosynthetic organs, is consistent with these interpretations. Thus, efficient light receiving of GP9-7 attributed to the distinctive plant shape during the period from 50 to 57 DAP that could be associated with the temporary improvement of the NAR during this period (**Figure 3D**). In the subsequent period (57 to 85 DAP), GP9-7 had the extremely higher LAI, which raises the risk of mutual shielding as an inevitable consequence. However, the plant shape transition of GP9-7 from 57 DAP, when tillers began to erect (**Figure 2A**), could adaptively sustain a preferred light receiving character in the thicker stand.

Weed Control in GP9-7 Stand

Ground coverage proceeded more rapidly in GP9-7 than in Koshihikari (**Figure 4A**), a factor that restricted light penetration onto the ground surface. Furthermore, the RPPFD just above the ground surface under the GP9-7 canopy at 53 DAP was 5% of that above the canopy and significantly lower than that of Koshihikari (**Figure 4B**). Since the lower ground level RPPFD is preferable for optimal weed control (Gibson et al., 2003; Koarai and Morita, 2003), the RPPFD of the ground level in GP9-7 stands was examined in detail in Fukuyama, a region in

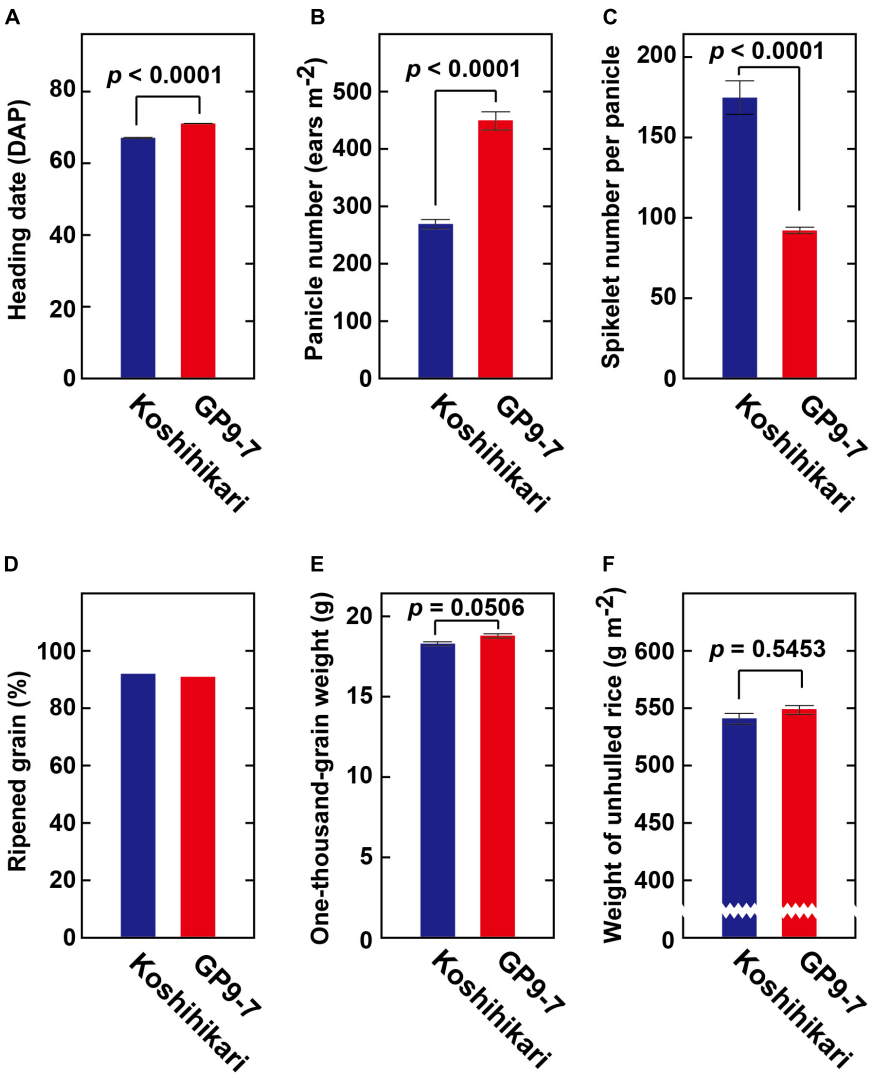


FIGURE 7 | Yield-related traits of GP9-7. **(A)** Heading date (n = 12), **(B)** panicle number per m^2 (n = 12), **(C)** spikelet number per panicle (n = 12), **(D)** percent ripened grain (n = 5), **(E)** one-thousand grain weight (n = 3), and **(F)** weight of unhulled rice (n = 3) of Koshihikari and GP9-7 cultivated in the test field at Tsukuba, Japan in the summer of 2018 (planting date: May 28, 2018) under sparse conditions (interplant space = 36 cm; 7.7 plants m^{-2}). P-values were calculated by the Welch's t-test and are reported above the bar graphs. Mean values for Koshihikari and GP9-7 with SE values are indicated with blue and red bars, respectively.

western Japan. In this study, the planting date was mid-June (June 15, 2019), about a month later than that at Tsukuba (mid-May). Although the heading date relative to the DAP shown in **Figure 5** was significantly different from experiments conducted in Tsukuba (**Figures 2–4**), the progression of vegetative growth was essentially similar to that of Tsukuba-grown plants. After planting, the number of GP9-7 tillers increased (**Figure 5A**) and spread uniformly, significantly decreasing RPPFD of the ground level at approximately 30 DAP (**Figure 5A**). After 35 DAP, Koshihikari and GP9-7 canopies covered the soil as shown by their similar and continued low RPPFD values. Tiller emergence of the transplanted weed, *E. crus-galli* var. *formosensis*, one of the most serious weeds in Japan, was significantly suppressed in GP9-7 stand (**Figure 5B**). The preceding attenuation of incident light under the GP9-7 canopy (**Figure 5A**) acted to suppress initial weed growth to a level insufficient to increase weed spread. Also, GP9-7 exhibited remarkable weed suppressive activity against the naturally occurring weed, *E. crus-galli* var. *formosensis* (**Figure 6**). Especially, tiller numbers and dry matter weights of naturally emerging weeds were significantly repressed under GP9-7 canopies (**Figures 6B,C**). Although GP9-7 did not completely eradicate the weeds in our experiment, the significant repression of weed growth under GP9-7 canopy may reduce nutrients interception by weeds and inhibit weed seed production, two factors that should help relieve farmers of weeding costs. Reduced numbers of herbicide applications will also decrease the environmental load and the risk arising herbicide-resistant weeds (Annett et al., 2014; Heap and Duke, 2018; Islam et al., 2018).

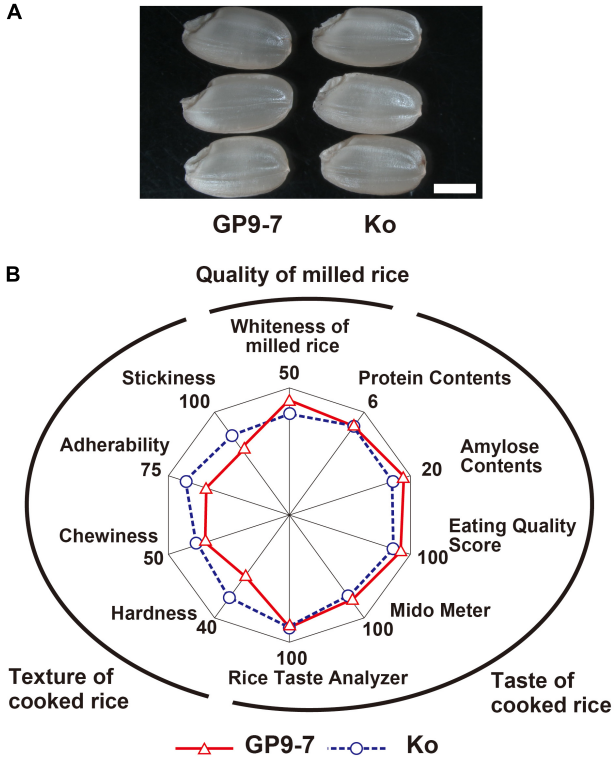


FIGURE 8 | Grain quality-related traits of GP9-7 harvested in the Tsukuba field in the fall of 2018. **(A)** A morphological comparison of typical hulled rice grains of Koshihikari and GP9-7. Scale bar: 2 mm. **(B)** Summary of rice texture and taste quality tests. Values for Koshihikari are indicated with blue circles with broken lines and values for GP9-7 are represented with red triangles with solid lines.

Yield and Grain Quality of GP9-7

GP9-7 was constructed using the genetic background of Koshihikari, the most widely accepted *japonica* cultivar for staple food in Japan due to its good taste. Since GP9-7 contains a small segment of chromosome 7 from *O. rufipogon* (**Figure 1A**), GP9-7 displayed slightly delayed heading (**Figure 7A**), almost twice the number of panicles per m^2 (**Figure 7B**), almost half the number of spikelets per panicles (**Figure 7C**), and a nearly equivalent amount of ripened grain (**Figure 7D**). The thousand grain weight of GP9-7 was also equivalent to that of Koshihikari (**Figure 7E**). Consequently, the final yield of GP9-7 was about the same as Koshihikari (**Figure 7F**). Significant increase of panicle number per plant (**Figure 7B**) should compensate the panicle number per area, which increase is restricted in existing cultivars under the sparse planting condition. On the other hand, significant promotion of GP9-7 vegetative growth (**Figure 3A**) did not result in an increase in yield (**Figure 7F**). We conducted yield tests on the Tsukuba fields from 2016 to 2020 and in the Fukuyama field in 2018 and 2019 (**Supplementary Table 3**). Despite the cultivation trials over the several summers, no significant increase in yield was detected. Possibly optimal cultivation conditions that bring out the best performance from GP9-7 have not been identified yet. Although the source ability of rice was improved in this study, the sink ability was not manipulated. This possibility may be an alternative reason for why the yield of the NIL did not increase in our experiment. Nevertheless, the introduction of a chromosomal segment that promoted vegetative growth and comparable grain yields is a significant advance since the yields of inbred lines to which a similar chromosomal segment was introduced were reported to be halved (Tan et al., 2008; Hua et al., 2016; Wu et al., 2018). Furthermore, in the yield tests conducted on the Fukuyama fields, no significant reduction of the grain yields of GP9-7 compared with those of Koshihikari was detected even in the semi-dense planting density (**Supplementary Table 3**), which cultivation condition is close to the previous reports (Tan et al., 2008; Hua et al., 2016; Wu et al., 2018).

A similar inbred line, YIL18, was derived from a cross between the *indica* cultivar, Teqing, as the recipient parent and *O. rufipogon* (accession: YJCWR), as the donor parent (Tan et al., 2008). Another similar inbred line, DIL29, was derived from a cross between the *indica* cultivar, Guichao 2, as the recipient parent and *O. rufipogon* (accession: DXCWR), as the donor parent (Wu et al., 2018). Both the donor accessions, unlike ours, were collected in China. The grain yields of YIL18 and DIL29 were about 57% of Teqing and about 63% of Guichao 2, significantly lead to inferior yields (Tan et al., 2008; Hua et al., 2016; Wu et al., 2018). Hua et al. (2016) examined the canopy structure of YIL18 using three-dimensional digitizing analysis and found a significant decrease in the LAI of YIL18 compared with those of the recurrent parent, Teqing. This report also

mentioned that YIL18 had greater mutual shading due to its prostrate plant shape.

We attribute this major discrepancy to three reasons: (1) Inbred lines YIL18 and DIL29 contain several wild rice chromosome segments in addition to the chromosome 7 segment. These additional chromosome segments may be responsible for the yield reduction; (2) In these cases, the recurrent parents were the *indica* cultivars, whereas we used Koshihikari, a *japonica* cultivar; and (3) The chromosomal donor in this study was collected from Thailand, whereas the donors in previous studies were collected from China. It is possible that differences in the sequence of this chromosomal segment may impart different traits. In fact, the *Prog1* sequence of GP9-7 was different from those of accessions YJCWR and DXCWR (**Supplementary Figure 3B**). Progress in whole-genome sequence analysis of these accessions may shed light on the reasons for these discrepancies. Identifying which of the three types of *O. rufipogon* (Or-I, -II, and -III) (Huang et al., 2012) by genome-wide association analyses that accurately describe these accessions may sort out these confusing results.

The surface of GP9-7 milled grains was the white-like Koshihikari (**Figure 8**). The taste-related factors of GP9-7 were nearly identical except for the amylose content that was very slightly higher in GP9-7 compared to Koshihikari (**Figure 8B**). Cooked rice from GP9-7 had relatively softer texture factors (**Figure 8B**), suggesting that the introduced chromosomal segment of wild rice in this study did not negatively influence grain quality, including its taste.

The remaining flaw of GP9-7 was lowered lodging resistance attributed to the smaller stem diameters that are a trade-off for the increased number of tillers. We believe that improvements such as the introduction of alleles conferring lodging resistance are possible. Lodging resistance loci have been well studied and exploitation of them will overcome this flaw (Shah et al., 2019).

Influences of This Research on the Future

This study attained that construction of the NIL capable of accelerating vegetative growth and suppressing weeds using wild rice genetic resources without undesirable consequences and yield reduction. The practical use of the NIL or the chromosomal segment that causes these traits may reduce weed control costs and environmental consequences associated with herbicide application, which open a way to solve some obstacles to the United Nations SDGs. This study also indicates that the genetic diversity of wild rice deserted in the domestication process is a valuable resource for breeding new cultivars with desired traits. The Japanese concept of *MOTTAINAI* has to be pushed to the front in this case; we have an obligation to excavate and apply these precious resources hidden in wild relatives in order to realize a bright 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: GenBank/EMBL/DDBJ Acc. Nos. LC573903 and LC573904.

AUTHOR CONTRIBUTIONS

NI, HA, and KI designed the research and performed the experiments. HH contributed to the construction of the NIL. AU and TI provided critical supports for assaying weed control. NI and HA wrote most of the manuscript with help from all the other authors. All authors read and approved the manuscript.

FUNDING

This study was supported by grants from commissioned project studies on the "Development of Labor-Saving Management of Serious Weeds to Expand Cultivation of Direct-Seeded Rice (19190995)" and "Genomics-Based Technology for Agricultural Improvement (RBS-2010)," Ministry of Agriculture, Forestry and Fisheries, Japan.

ACKNOWLEDGMENTS

The authors thank Tachibana (Western Region Agricultural Research Center/NARO), Koarai (Institute for Plant Protection/NARO), and Sentoku (Institute of Agrobiological Sciences/NARO) for supplying their informative insights and knowledge.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.748531/full#supplementary-material

REFERENCES

Agarwal, P., Arora, R., Ray, S., Singh, A. K., Singh, V. P., Takatsuji, H., et al. (2007). Genome-wide identification of C2H2 zinc-finger gene family in rice and their phylogeny and expression analysis. *Plant Mol. Biol.* 65, 467–485. doi: 10.1007/s11103-007-9199-y

Annett, R., Habibi, H. R., and Hontela, A. (2014). Impact of glyphosate and glyphosate-based herbicides on the freshwater environment. *J. Appl. Toxicol.* 34, 458–479. doi: 10.1002/jat.2997

Chauhan, B. S. (2020). Grand challenges in weed management. *Front. Agron.* 1:3. doi: 10.3389/fagro.2019.00003

Chen, E., Huang, X., Tian, Z., Wing, R. A., and Han, B. (2019). The genomics of *Oryza* species provides insights into rice domestication and heterosis. *Annu. Rev. Plant Biol.* 70, 639–665. doi: 10.1146/annurev-arplant-050718-100320

Dass, A., Shekhawat, K., Choudhary, A. K., Sepat, S., Rathore, S. S., Mahajan, G., et al. (2017). Weed management in rice using crop competition-a review. *Crop Prot.* 95, 45–52. doi: 10.1016/j.cropro.2016.08.005

Doebley, J. F., Gaut, B. S., and Smith, B. D. (2006). The molecular genetics of crop domestication. *Cell* 127, 1309–1321. doi: 10.1016/j.cell.2006.12.006

Duncan, W. G., Loomis, R. S., Williams, W. A., and Hanau, R. (1967). A model for simulating photosynthesis in plant communities. *Hilgardia* 38, 181–205. doi: 10.3733/hilg.v38n04p181

Fei, C., Yu, J., Xu, Z., and Xu, Q. (2019). Erect panicle architecture contributes to increased rice production through the improvement of canopy structure. *Mol. Breed.* 39:128. doi: 10.1007/s11032-019-1037-9

Fischer, A. J., Ramírez, H. V., Gibson, K. D., Da Silveira, and Pinheiro, B. (2001). Competitiveness of semidwarf upland rice cultivars against palisadegrass (*Brachiaria brizantha*) and signalgrass (*B. decumbens*). *Agron. J.* 93, 967–973. doi: 10.2134/agronj2001.935967x

Gibson, K. D., Fischer, A. J., Foin, T. C., and Hill, J. E. (2003). Crop traits related to weed suppression in water-seeded rice (*Oryza sativa* L). *Weed Sci.* 51, 87–93.

Global Rice Science Partnership (2013). *Rice Almanac*, 4th Edn. Los Banþos: International Rice Research Institute.

Guo, Z.-W., Deng, H.-F., Li, S.-Y., Xiao, L.-T., Huang, Z.-Y., et al. (2011). Characteristics of the mesophyllous cells in the sheaths of rice (*Oryza sativa* L.). *Agri. Sci. China* 10, 1354–1364. doi: 10.1016/S1671-2927(11)60128-4

Heap, I., and Duke, S. O. (2018). Overview of glyphosate-resistant weeds worldwide. *Pest Manag. Sci.* 74, 1040–1049. doi: 10.1002/ps.4760

Hirabayashi, H., Sato, H., Nonoue, Y., Kuno-Takemoto, Y., Takeuchi, Y., Kato, H., et al. (2010). Development of introgression lines derived from *Oryza rufipogon* and *O. glumaepatula* in the genetic background of *japonica* cultivated rice (*O. sativa* L.) and evaluation of resistance to rice blast. *Breed. Sci.* 60, 604–612. doi: 10.1270/jsbbs.60.604

Hua, S., Cao, B., Zheng, B., Li, B., and Sun, C. (2016). Quantitative evaluation of influence of *PROSTRATE GROWTH 1* gene on rice canopy structure based on three-dimensional structure model. *Field Crops Res.* 194, 65–74. doi: 10.1016/j.fcr.2016.05.004

Huang, L., Liu, H., Wu, J., Zhao, R., Li, Y., Melaku, G., et al. (2020). Evolution of plant architecture in *Oryza* driven by the *PROG1* locus. *Front. Plant Sci.* 11:876. doi: 10.3389/fpls.2020.00876

Huang, X., Kurata, N., Wei, X., Wang, Z. X., Wang, A., Zhao, Q., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501. doi: 10.1038/nature11532

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895

Islam, F., Wang, J., Farooq, M. A., Khan, M. S. S., Xu, L., Zhu, J., et al. (2018). Potential impact of the herbicide 2,4-dichlorophenoxyacetic acid on human and ecosystems. *Environ. Int.* 111, 332–351. doi: 10.1016/j.envint.2017.10.020

Jin, J., Huang, W., Gao, J. P., Yang, J., Shi, M., Zhu, M. Z., et al. (2008). Genetic control of rice plant architecture under domestication. *Nat. Genet.* 40, 1365–1369. doi: 10.1038/ng.247

Johnson, D. E., Dingkuhn, M., Jones, M. P., and Mahamane, M. C. (1998). The influence of rice plant type on the erect of weed competition on *Oryza sativa* and *Oryza glaberrima*. *Weed Res.* 38, 207–216. doi: 10.1046/j.1365-3180.1998.00092.x

Kagale, S., Links, M. G., and Rozwadowski, K. (2010). Genome-wide analysis of ethylene-responsive element binding factor-associated amphiphilic repression motif-containing transcriptional regulators in Arabidopsis. *Plant Physiol.* 152, 1109–1134. doi: 10.1104/pp.109.151704

Kamboj, R., Singh, B., Mondal, T. K., and Bisht, D. S. (2020). Current status of genomic resources on wild relatives of rice. *Breed. Sci.* 70, 135–144. doi: 10.1270/jsbbs.19064

Koarai, A., and Morita, H. (2003). Evaluation of the suppression ability of rice (*Oryza sativa*) on *Monochoria vaginalis* by measuring photosynthetic photon flux density below rice canopy. *Weed Biol. Manag.* 3, 172–178. doi: 10.1046/j.1445-6664.2003.00104.x

McCouch, S. R., Teytelman, L., Xu, Y., Lobos, K. B., Clare, K., Walton, M., et al. (2002). Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res.* 9, 199–207. doi: 10.1093/dnares/9.6.199

Monteith, J. L. (1965). Light distribution and photosynthesis in field crops. *Annal. Bot.* 29, 17–37. doi: 10.1093/oxfordjournals.aob.a083934

Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Philos. Trans. Royal Soc. B* 281, 277–294. doi: 10.1098/rstb.1977.0140

Sakamoto, T., Morinaka, Y., Ohnishi, T., Sunohara, H., Fujioka, S., Ueguchi-Tanaka, M., et al. (2006). Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. *Nat. Biotechnol.* 24, 105–109. doi: 10.1038/nbt1173

San, N. S., Ootsuki, Y., Adachi, S., Yamamoto, T., Ueda, T., Tanabata, T., et al. (2018). A near-isogenic rice line carrying a QTL for larger leaf inclination angle yields heavier biomass and grain. *Field Crops Res.* 219, 131–138. doi: 10.1016/j.fcr.2018.01.025

Shah, L., Yahya, M., Shah, S. M. A., Nadeem, M., Ali, A., Ali, A., et al. (2019). Improving lodging resistance: using wheat and rice as classical examples. *Int. J. Mol. Sci.* 20:17. doi: 10.3390/ijms20174211

Tan, L., Li, X., Liu, F., Sun, X., Li, C., Zhu, Z., et al. (2008). Control of a key transition from prostrate to erect growth in rice domestication. *Nat. Genet.* 40, 1360–1364. doi: 10.1038/ng.197

Temnykh, S., Park, W. D., Ayres, N., Cartinhour, S., Hauck, N., Lipovich, L., et al. (2000). Mapping and genome organization of microsatellite sequences in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100, 697–712. doi: 10.1007/s001220051342

Wu, Y., Zhao, S., Li, X., Zhang, B., Jiang, L., Tang, Y., et al. (2018). Deletions linked to *PROG1* gene participate in plant architecture domestication in Asian and African rice. *Nat. Commun.* 9:4157. doi: 10.1038/s41467-018-06509-2

Yu, B., Lin, Z., Li, H., Li, X., Li, J., Wang, Y., et al. (2007). *Tac1*, a major quantitative trait locus controlling tiller angle in rice. *Plant J.* 52, 891–898. doi: 10.1111/j.1365-313X.2007.03284.x

Zhu, Q., Zheng, X., Luo, J., Gaut, B. S., and Ge, S. (2007). Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol. Biol. Evol.* 24, 875–888. doi: 10.1093/molbev/msm005

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 January 2022
doi: 10.3389/fgene.2021.802738

A High-Density Genetic Map Enables Genome Synteny and QTL Mapping of Vegetative Growth and Leaf Traits in Gardenia

Yang Cui[1], *Baolian Fan*[1], *Xu Xu*[1], *Shasha Sheng*[1], *Yuhui Xu*[2]* *and Xiaoyun Wang*[1]*

[1]Research Center for Traditional Chinese Medicine Resources and Ethnic Minority Medicine, Jiangxi University of Chinese Medicine, Nanchang, China, [2]Adsen Biotechnology Co., Ltd., Urumchi, China

OPEN ACCESS

Edited by:
Jinyoung Y. Barnaby, Agricultural Research Service (USDA), United States

Reviewed by:
Chengsong Zhu, University of Texas Southwestern Medical Center, United States
Xingbo Wu, Agricultural Research Service (USDA), United States

***Correspondence:**
Yuhui Xu
genetics_2010@163.com
Xiaoyun Wang
wxy20052002@aliyun.com

Specialty section:
This article was submitted to Plant Genomics, a section of the journal Frontiers in Genetics

Received: *27 October 2021*
Accepted: *13 December 2021*
Published: *04 January 2022*

Citation:
Cui Y, Fan B, Xu X, Sheng S, Xu Y and Wang X (2022) A High-Density Genetic Map Enables Genome Synteny and QTL Mapping of Vegetative Growth and Leaf Traits in Gardenia. Front. Genet. 12:802738. doi: 10.3389/fgene.2021.802738

The gardenia is a traditional medicinal horticultural plant in China, but its molecular genetic research has been largely hysteretic. Here, we constructed an F_1 population with 200 true hybrid individuals. Using the genotyping-by-sequencing method, a high-density sex-average genetic map was generated that contained 4,249 SNPs with a total length of 1956.28 cM and an average genetic distance of 0.46 cM. We developed 17 SNP-based Kompetitive Allele-Specific PCR markers and found that 15 SNPs were successfully genotyped, of which 13 single-nucleotide polymorphism genotypings of 96 F_1 individuals showed genotypes consistent with GBS-mined genotypes. A genomic collinearity analysis between gardenia and the *Rubiaceae* species *Coffea arabica*, *Coffea canephora* and *Ophiorrhiza pumila* showed the relativity strong conservation of LG11 with NC_039,919.1, HG974438.1 and Bliw01000011.1, respectively. Lastly, a quantitative trait loci analysis at three phenotyping time points (2019, 2020, and 2021) yielded 18 QTLs for growth-related traits and 31 QTLs for leaf-related traits, of which *qBSBN7-1*, *qCD8* and *qLNP2-1* could be repeatably detected. Five QTL regions (*qCD8* and *qSBD8*, *qBSBN7* and *qSI7*, *qCD4-1* and *qLLLS4*, *qLNP10* and *qSLWS10-2*, *qSBD10* and *qLLLS10*) with potential pleiotropic effects were also observed. This study provides novel insight into molecular genetic research and could be helpful for further gene cloning and marker-assisted selection for early growth and development traits in the gardenia.

Keywords: genetic map, genotyping-by-sequencing, growth-and leaf-related traits, QTL, synteny, gardenia

Abbreviations: SNPs, single-nucleotide polymorphisms; QTLs, quantitative trait loci; LOD, logarithm of odds; MAS, marker-assisted selection; NGS, next-generation sequencing; GBS, genotyping-by-sequencing; CD, crown diameter; BSBN, basal stem branch number; SI, stem inclination; PH, plant height; MSH, main stem height; SBD, stem base diameter; LNS, leaf number on stem; LNP, leaf number per plant; LLLS, longest leaf length on stem; LLWS, longest leaf width on stem; SLLS, shortest leaf length on stem; SLWS, shortest leaf width on stem; CV, coefficient of variation; LGs, linkage groups; PVE, phenotypic variance explained; KASP, kompetitive allele-specific PCR; CP, cross pollination; DEGs, differentially expressed genes.

INTRODUCTION

Gardenia (*Gardenia jasminoides* Ellis, 2n = 22) originated in central China, and it is a perennial shrub in the Rubiaceae family with edible flowers and medicinal fruits. Its dried ripe fruit has high quantities of crocin, geniposide, and genipin compounds (Chen Q. et al., 2020) and therefore possesses anti-inflammatory, antidepressant, anti-diabetes, antioxidant and antihypertensive activities (Qin et al., 2013; Higashino et al., 2014; Khajeh et al., 2020). The fruits are used in many traditional Chinese medicine preparations and formulas to treat different diseases (Chen L. et al., 2020). In addition to applications in traditional Chinese medicine, extracts of gardenia fruit are used as a natural colorant in the food and textile industries (Chen L. et al., 2020). Gardenia has beautiful fragrant flowers and evergreen leaves, so it is widely used for garden decoration. Fresh flowers are also used in China as edible vegetables or used to extract essential oils (Wang et al., 2017). Gardenia has a cultivation history of more than 1,000 years in China and was gradually introduced to Africa, Asia, Australia, Europe, North and South America, and the Pacific islands because of its medicinal, ornamental and industrial value (Xu et al., 2020).

Using traditional phenotypic selection-based breeding methods for genetic improvement is a labor- and time-consuming process because of the long lifecycle and highly heterozygous nature of the gardenia. By contrast, marker-assisted selection (MAS) using tightly linked or functional molecular markers with elite traits is an ideal approach to improving breeding efficiency (Mathew et al., 2014; Pootakham et al., 2015; Dong et al., 2019). However, the current molecular biology research for gardenia falls further behind model species, primarily focusing on phenotype, genetic evaluation or accession discrimination (Tsanakas et al., 2013; Hu et al., 2019; Wei et al., 2019; Li et al., 2021). Very limited studies on molecular marker identification in gardenia have been reported, such as dozens of SSR developments (Xu et al., 2014; Deng et al., 2015). Recently, a chromosomal-level genome assembly for the gardenia was released to dissect the pathway of crocin biosynthesis (Xu et al., 2020). Furthermore, helix-loop-helix (bHLH) transcription factors responsible for crocin biosynthesis were identified based on the gardenia genome (Tian et al., 2020). Genome assembly will undoubtedly accelerate functional genomics studies in gardenia. Nevertheless, the notably shortage of genome-wide molecular marker and the large gap between the phenotyping and genotyping are still bottlenecks for gardenia genetic improvement by molecular breeding and thus restrict the gardenia related industry.

Genetic maps based on the F_1 segregating population are a robust tool for identifying the linkage between traits and molecular markers, which have long been applied widely in highly heterozygous species of trees, flowering plants and aquatics (Jorge et al., 2005; Wang et al., 2006; Lambert et al., 2007; Oyant et al., 2007; Sánchez-Pérez et al., 2012; Pacheco et al., 2014). In the next-generation sequencing (NGS) era, sequencing-based technologies can provide novel strategies for genome-wide SNP (single-nucleotide polymorphism) development and help to construct a high-density genetic linkage map for high-resolution QTL (quantitative trait loci) identification (Rehman et al., 2020). SNP markers can be named in many ways, including reduced-representation sequencing, resequencing and transcriptome sequencing. Reduced-representation sequencing has been differentiated into different technologies, including genotyping-by-sequencing (GBS), restriction site-associated DNA sequencing (RAD-Seq), double-digest RAD (ddRAD), specific-locus amplified fragment sequencing (SLAF-seq), ezRAD (Toonen et al., 2013) and 2b-restriction site-associated DNA sequencing (2b-RAD) (Baird et al., 2008; Elshire et al., 2011; Peterson et al., 2012; Wang et al., 2012; Sun et al., 2013; Toonen et al., 2013). Notably, GBS is a feasible SNP discovery method for highly diverse and large genome species, even without reference genome, and it has been widely adopted in genotyping for genetic map construction (İpek et al., 2017; Gabay et al., 2018; Paudel et al., 2018; Robledo et al., 2018; Lewter et al., 2019; Rubio et al., 2020). For instance, a high-density linkage map of coffee, a tree belongs to the same *Rubiaceae* family with gardenia, was constructed using 848 SSR and SNP markers, of which the SNP markers were developed by GBS (Moncada et al., 2016). Additional high-density genetic maps with 3,000–6,000 SNP markers have been reported in many perennial plants (Pootakham et al., 2015; Zhang et al., 2016; İpek et al., 2016; Tello et al., 2019; Zhang et al., 2019). GBS was also used for genetic diversity analysis in coffee (Anagbogu et al., 2019).

Genetic map can provide chromosome-level variation information across species. In fact, the microsynteny and macrosynteny relationship have long been verified in plants (Paterson et al., 2004; Yan et al., 2004). Comparative mapping can illustrate the co-located molecular marker distribution patterns between different genome of organisms, and further reveal structural variations and collinearity among chromosomes. Using this method, a high degree of colinearity and chromosome recombination and inversion has been found in *Salicaceae* species (Hanley et al., 2006; Berlin et al., 2010). Similarly, chromosomal translocations and inversions were confirmed by comparing an eggplant genetic map with the genome sequence of both tomato and pepper (Rinaldi et al., 2016). Lately, the genomic evolutionary of *Coffea canephora* and *Ophiorrhiza pumila* were investigated (Zhao et al., 2021), and some high collinearity pairs and potential karyotype rearrangement were observed, indicating their chromosomal evolution in genomic differentiation (Kai et al., 2011; Kodama et al., 2014).

QTL mapping is a traditional method to build an association bridge between genotypes and phenotypes. The tightly linked markers in QTL regions can potentially be used for MAS (Chang et al., 2018; Kim et al., 2018; Yamakawa et al., 2021). The phenotypes for typical QTL mapping always focus on specific developmental stages, and the identified QTLs represent the accumulation effect of related gene expression at the phenotyping stages. However, plant growth and development are dynamic, ever-changing processes. Dynamic QTL analysis enables QTL detection for target traits over the entire developmental process, especially for tree species, which require a relatively long time for morphogenesis. Dynamic QTL mapping studies have been published primarily for crops

such as rice (Sun et al., 2015), maize (Wang et al., 2019), wheat (Mohler and Stadlmeier, 2019), cotton (Shang et al., 2015) and oilseed rape (Wang et al., 2015). In tree species, however, limited dynamic QTL maps were conducted. Desnoues et al. (2016) reported the dynamic QTL mapping of fresh weight, sugar, acid and enzyme activity at different developmental stages of peach fruit, and observed the effect of allele changes during fruit ripening. Recently, the leaf traits and plant height of *Catalpa bungei* at five successive time points were investigated, and a total of 33 QTLs were mapped using a high-density genetic map (Lu et al., 2019). In *Populus*, a total of 311 QTLs for three growth traits at 12 time points were mapped, and many QTLs specific to one time point were identified (Du et al., 2019). These results illustrated the importance of dynamic QTL mapping for the genetic dissection of developmental traits.

The genetic map of the gardenia has not been released to date. In the present study, we used a paternity test-passed F_1 population of gardenia, and then employed GBS technology to construct a high-density genetic map for collinearity analysis between *Rubiaceae* species. Moreover, a high-resolution dynamic QTL mapping analysis was performed on growth and leaf related traits during the vegetative growth stage for three continuous years. This study was the first high-density genetic map-based QTL study in gardenia, laying a foundation for further gene cloning and MAS breeding.

MATERIALS AND METHODS

Mapping Population Construction and Phenotyping

We previously screened two *Gardenia jasminoides* Ellis. germplasms that exhibited distinct phenotypes, namely, GD1 with high branches, large fruit, medium leaf widths and a broad crown type and AX5 with dwarf branches, small fruit, narrow leaf widths and a thin crown type. In May 2017, following emasculation at the early stage of flower development, GD1 (♀) and AX5 (♂) were crossed by artificial pollination. The dark red fruits were harvested during the first frost. The hybrid seeds were isolated and then placed on moist germination paper in Petri dishes in November 2017. At the time of radicle protrusion, the seeds were transferred into pots in the greenhouse. During the following year on March 27, the seedlings were transplanted within the Botanic Garden at Jiangxi University of Chinese Medicine (N28N°40′, E115°45′). The two parents and a total of 207 F_1 individuals were randomly planted. In April 2019, young leaves from the two parents and all the F_1 individuals were harvested and stored in a silica-gel drier for further DNA extraction.

Phenotyping and Data Processing

We measured 12 traits over three continuous years in October 2019, July 2020 and April 2021, and all the traits were measured three times. The detailed measurements are shown in **Table 1** and **Figure 1**. Protractors and Vernier calipers were used to measure the stem inclinations and stem base diameters, respectively. The flexible rule was used to measure the remaining traits. SPSS V17.0 software (SPSS Inc., Chicago, IL, United States) was used for variance analysis. TBtools was used to display the variation and Pearson pairwise correlations graphically among different traits (Chen C. et al., 2020).

Paternity Test

To ensure an expectant hybrid seed panel, a paternity test was conducted using simple sequence repeat (SSR) markers. The total DNA from the two parents and 207 F_1 individuals was isolated separately using DNA Rapid Extraction Kit DN1403 (Aidlab Biotechnologies Co., Ltd., Beijing, China). Using the total DNA from the two parents, a total of 25 SSRs from Deng et al. (2015) were used for polymorphic screening. Homozygous and polymorphic SSR markers were selected to genotype the 207 F_1 individuals. For example, if the genotypes of the two parents were encoded with "aa" and "bb", then the genotype of the true F_1 offspring was "ab". The polymerase chain reaction (PCR) system for SSR genotyping was performed in a 10.0 μl volume, with 5 μl 2×Taq MasterMix, 0.2 μl forward primer (F) and reverse primer (R), respectively 1 μl sample genomic DNA and 3.6 μl ddH_2O. The system was pre-degenerated at 94°C for 3 min, and then PCR amplification began for 34 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 30 s, and a final extension at 72°C for 5 min. An 8% denaturing polyacrylamide gel was used to separate the PCR products for further silver staining.

Population GBS Sequencing and Genotyping

Similar to the paternity test, genomic DNA was isolated from the young leaves of GD1, AX5 and 200 true hybrid F_1 individuals using the DNA Rapid Extraction Kit DN1403 (Aidlab Biotechnologies Co., Ltd., Beijing, China). The DNA concentration and quality were monitored using a NanoDrop spectrophotometer (ND 2000, Thermo Fisher Scientific, United States) and electrophoresis on 0.85% agarose gels, respectively. Then, GBS libraries were constructed. In brief, the genomic DNA was placed into a combination solution of *RsaI* and *HaeIII* for digestion. Products between 429 and 459 bp in length were enriched in 3% agarose gels, and end repair was performed with End Prep Enzyme Mix, followed by 3′A extension and adaptor addition. The dual index for further sample identification was introduced by PCR with eight cycles. Library quantification was performed using an Agilent 2,100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States), and all the libraries were mixed into one lane for paired-end sequencing (PE150) at Adsen Biotechnology Co., Ltd (Urumchi, China) using an Illumina NovaSeq 6,000 (Illumina, San Diego, CA, United States). The raw data were filtered to generate high-quality clean data according to Zhao's criteria (Zhao et al., 2021). The genotyping was processed according to the following steps. First, a Burrows-Wheeler aligner (Li and Durbin, 2009) was used to map the clean reads to the reference genome of gardenia (Xu et al., 2020), followed by duplicate removal (Picard: http://sourceforge.net/projects/picard/). Second, SNPs were called by combining the HaplotypeCaller module of GATK (McKenna et al., 2010) and SAMtools (Li et al., 2009) to guarantee a high-quality SNP dataset. Lastly, dual-detected SNPs with sequencing depths ≥8 in the two parents, segregation distortion $p > 0.01$ (Chi-square) and integrity ≥60% in the offspring were maintained and encoded into

TABLE 1 | Detailed measurement methods for the 12 agronomic traits.

Trait	Abbreviation	Description
Crown diameter	CD	Measuring the diameter of the identifiable three-dimensional cylinder of each individual tree
Basal stem branch number	BSBN	Counting the branch numbers derived from the basal stem
Stem inclination	SI	See **Figure 1**
Plant height	PH	See **Figure 1**
Main stem height	MSH	See **Figure 1**
Stem base diameter	SBD	Diameter of the stem base
Leaf number on stem	LNS	Counting all the leaf numbers on the main stem
Leaf number per plant	LNP	Counting all the leaf numbers per plant
Longest leaf length on stem	LLLS	Length of the longest leaf on the stem **(Figure 1)**
Longest leaf width on stem	LLWS	Width of the longest leaf on the stem **(Figure 1)**
Shortest leaf length on stem	SLLS	Length of the shortest leaf on the stem **(Figure 1)**
Shortest leaf width on stem	SLWS	Width of the shortest leaf on the stem **(Figure 1)**

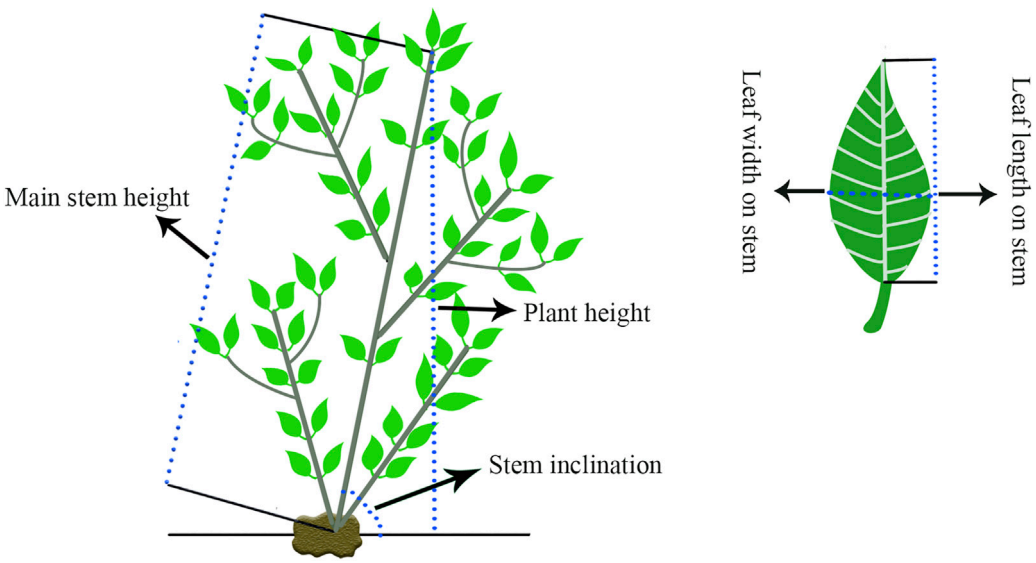


FIGURE 1 | Part of the trait measurement schematic diagram.

eight genotyping patterns suitable for diploid species (aa × bb, ab × cd, ef ×, e.g., hk × hk, lm × ll, nn × np, ab × cc, and cc × ab). All the genotypes except aa × bb were selected and SMOOTH algorithms (van Os et al., 2005) were used to correct genotypes and imputation for further genetic map construction.

Genetic Linkage Map Construction, QTL Mapping and Gene Annotation Analyses

All the retained SNP markers were assigned into linkage groups (LGs) based on the mapping location on the reference genome of the gardenia (Xu et al., 2020). JoinMap software (V4.1) was applied for linear arrangement within LGs using the mapping function of the cross pollination (CP) model (Van Ooijen 2006). Map distances were estimated using the Kosambi mapping function (Kosambi, 1943). Genetic map visualization was performed using a CheckMatrix heat plot (http://cgpdb.ucdavis.edu/XLinkage/). The Spearman correlation coefficient between the final LGs and the reference genome was calculated and visualized using R (www.r-project.org/). MapQTL V6.0 was used for QTL analyses using the interval mapping (IM) algorithm (Van Ooijen 2009). QTLs were cut off when the LOD (logarithm of odds) values of three continuous SNPs were ≥2.5. Genes underlying stable expressed QTLs were annotated by ANNOVAR (Wang et al., 2010), and functional enrichment analyses were conducted by UniProtKB/Swiss-Prot database (Schneider et al., 2004), Pfam (Bateman et al., 2004), Gene Ontology (Ashburner et al., 2000) and KEGG (kyoto encyclopedia of genes and genomes) (Kanehisa and Goto, 2000).

Genome Synteny Analyses

To explore the evolutionary relationship between gardenia and other *Rubiaceae* species with chromosome-level genomes, SNP-based high-density genetic maps were aligned to the genomes of *Coffea canephora* (https://www.ncbi.nlm.nih.gov/genome/12248), *Coffea arabica* (https://www.ncbi.nlm.nih.gov/genome/

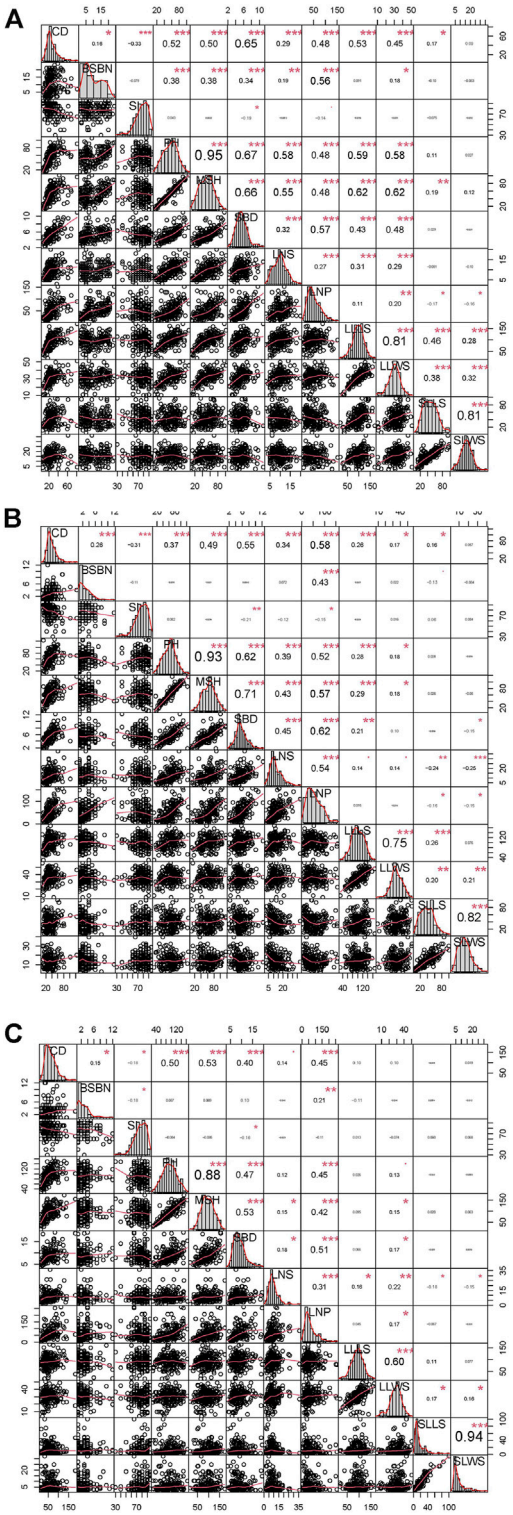


FIGURE 2 | Variation and Pearson pairwise correlation analyses of growth-related and leaf-related traits of the F_1 population. **(A)**, **(B)** and **(C)** represent the variation and Pearson pairwise correlations in 2019, 2020 and 2021, respectively. The correlations were calculated with Spearman correlation coefficients, and the *p* values are indicated as follows: *, $p < 0.05$; **, $p < 0.01$; and ***, $p < 0.001$. The abbreviations given in the histograms are (*Continued*)

?term=Coffea+arabica) and *Ophiorrhiza pumila* (https://www.ncbi.nlm.nih.gov/genome/97777?genome_assembly_id=1538555) using BLAST (Kent, 2002), and the physical positions of the homologous sequences were used to generate a collinearity diagram in R (www.r-project.org/).

SNP Confirmation by Kompetitive Allele-Specific PCR (KASP)

To confirm the SNPs developed by GBS, we randomly genotyped 96 F_1 individuals by KASP using 17 SNPs from five randomly selected QTL regions (**Supplementary Table S1**). Primer 5.0 was used to design the primers, and BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) was used to check the primer specificity. The primer information is shown in **Supplementary Table S1**. The KASP genotyping processes were conducted in the GeneMatrix system (HC Scientific, Chengdu, China) according to the following three parts: Matrix Arrayer reaction plate preparation apparatus, Matrix Cycler high-throughput water bath thermal cycler, and Matrix Scanner high-speed fluorescence scanner. The PCR system contained 1 μl 2×KASP Master mix (standard ROX) (LGC Biosearch Technologies, United KIngdom), 0.028 μl KASP primer mix and 1 μl sample DNA (~50 ng/μl). The detailed KASP thermal cycling program was 94°C for 15 min, followed by 10 cycles of 94°C for 20 s, 61–55°C for 20 s (dropping 0.6°C per cycle), 72°C for 45 s, 30 cycles of 94°C for 20 s, and 55°C for 1 min.

RESULTS

Hybridization Test

In the present study, the parents GD1 and AX5 were used as materials, and 47 published SSRs were used for polymorphism tests. A total of 19 pairs of primers were observed to be polymorphic. We further selected markers that were homozygous and polymorphic in the parents. That is, aa×bb-type polymorphic SSRs eGJ026 and eGJ118 were used for genotyping the progeny. A total of 200 progeny out of 207 individuals in the F_1 population were true hybrids with the segregation type "ab" (**Supplementary Table S1–S2**), indicating that the hybridization experiment was strictly controlled.

Genetic Variations in 12 Phenotypes

A set of relatively wide ranges of variations were observed in the crown diameter (CD), stem inclination (SI), plant height (PH), main stem height (MSH), leaf number per plant (LNP), the longest leaf length on stem (LLLS) and the shortest leaf length

FIGURE 2 | as follows: CD: crown diameter; BSBN: basal stem branch number; SI: stem inclination; PH: plant height; MSH: main stem height; SBD: stem base diameter; LNS: leaf number on stem; LNP: leaf number per plant; LLLS: longest leaf length on stem; LLWS: longest leaf width on stem; SLLS: shortest leaf length on stem; and SLWS: shortest leaf width on stem.

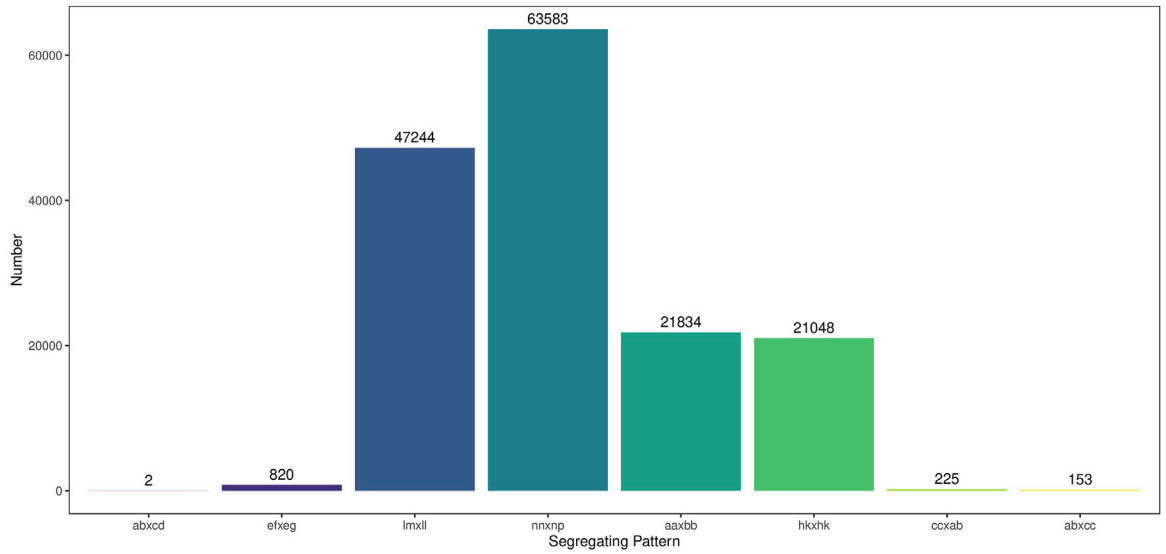


FIGURE 3 | The distributions of SNP marker segregation patterns.

TABLE 2 | The basic characteristics of the female genetic map, male genetic map and sex-average genetic map.

LG	Marker number			Gap≤ 5 cM (%)			Max gap (cM)			Total distance (cM)			Average distance (cM)		
	Sex-average	Female	Male	Sex-average	Female	Male	Sex-average	Female	Male	Sex-average	Female	Male	Sex-average	Female	Male
1	403	322	100	96.77	95.02	93.94	14.08	23.9	30.2	202.35	217.7	114.49	0.5	0.68	1.14
2	648	352	348	97.37	92.31	97.69	7.19	17.48	13.72	228.19	376.23	80.16	0.35	1.07	0.23
3	468	242	272	99.79	98.76	99.63	18.52	24.71	15.06	110.56	131.34	68.76	0.24	0.54	0.25
4	606	456	183	99.34	98.46	95.05	9.91	8.72	17.83	183.36	174.16	167.7	0.3	0.38	0.92
5	136	62	77	92.59	91.8	89.47	16.59	67.35	10.64	126.19	152.99	87.51	0.93	2.47	1.14
6	194	119	85	93.78	89.83	96.43	16.83	60.2	16.42	187.07	303.73	67.31	0.96	2.55	0.79
7	312	157	170	98.07	98.72	97.63	23.86	25.72	48.88	150.18	104.37	156.02	0.48	0.66	0.92
8	369	187	210	98.37	91.94	96.65	9.96	17.46	15.06	197.3	249.66	144.94	0.53	1.34	0.69
9	120	81	52	92.44	91.25	82.35	14.72	44.89	23.37	142.71	137.78	144.08	1.19	1.7	2.77
10	406	245	189	98.27	95.9	96.81	18.08	29.89	33.67	205.04	244.18	143.14	0.51	1	0.76
11	587	362	277	98.63	96.68	97.46	17.49	17.83	31.74	223.33	256.34	174.27	0.38	0.71	0.63
Total	4,249	2,585	1,963	96.86	94.61	94.83	23.86	67.35	48.88	1956.28	2,348.48	1,348.3	0.46	0.91	0.69

on stem (SLLS), while mild variations were present in the remaining five phenotypes (**Supplementary Table S2**). The coefficient of variation (CV) of the leaf-related trait LLLS and the longest leaf width on stem (LLWS) were primarily stationary at the three time points, which was similar to all six growth-related traits, suggesting minor differences among these three time points. Only the leaf-related traits LNS, LNP, SLLS and SLWS demonstrated acute CV fluctuation (**Supplementary Table S2**). Among every year for the phenotypes at the three time point, CD, PH and LNP exhibited strong correlations with other phenotypes. However, SI had no highly significant correlations with the other ten traits (except SBD). PH had the strongest positive correlations with MSH, with correlation coefficients of 0.95, 0.93 and 0.88 in 2019, 2020 and 2021, respectively. SLLS and SLWS also exhibited a correlation greater than 0.80 in the 3 years (**Figure 2**). The correlation analysis implied that there was an independent and interdependent relationship between growth-related traits and leaf-related traits.

Variation Calling and Genotyping

In total, GBS sequencing generated 29, 630, 679 clean reads after quality control, with 1.77 and 0.83 Gb for the parent AX5 and GD1, respectively. For the offspring, 12, 186, 237 reads (1.81 Gb) were obtained per individual. The statistics showed that the average Q30 was higher than 85%, and the GC content (%) was distributed between 40.29 and 48.18 (**Supplementary Table S3**). Upon using BWA software to align the sequencing data to the reference genome of gardenia, the mapping rates were 94.38, 93.88 and 96.66% for AX5, GD1 and all the progeny, respectively. These pre-processing procedures indicated a high quality of sequencing data for further analysis.

A total of 154,909 SNPs were detected by combining the GATK and SAMTools, and the genotypes of these SNPs were encoded into eight segregation patterns. Among them, lm×ll, np×nn, hk×hk and aa×bb occupied 47,244, 63,583, 21,048 and 21,834 SNPs, respectively, accounting for 99.23% of the total

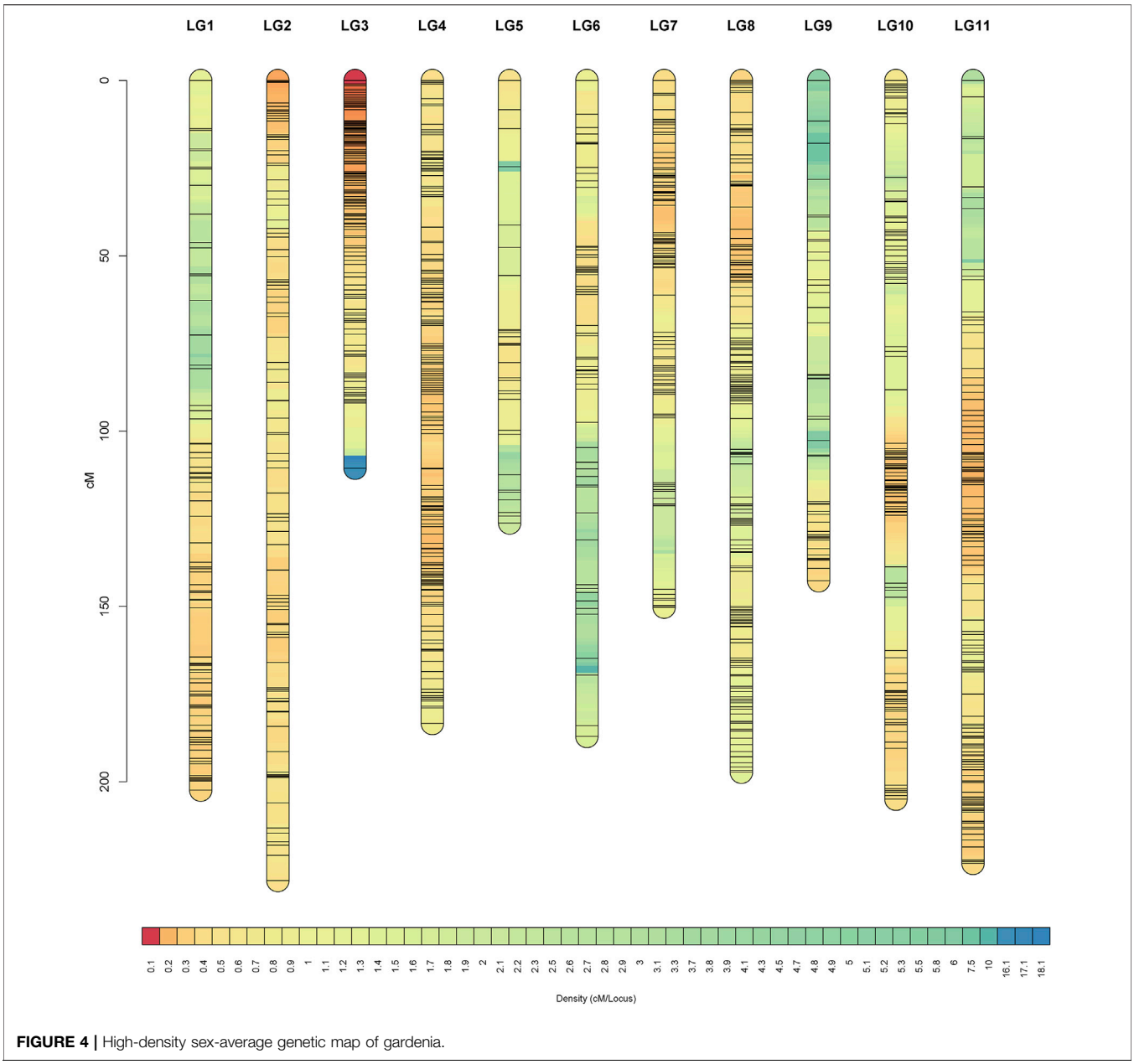


FIGURE 4 | High-density sex-average genetic map of gardenia.

SNPs (**Figure 3**). After depth and integrity filtering, the remaining SNPs were used for genetic map construction.

High Density Genetic Map

JoinMap was used to construct a female genetic map containing 2,585 markers spanning 2,348.48 cM and a male genetic map containing 1,963 markers spanning 1,348.38 cM, and both consisted of 11 LGs (**Table 2**). Integrating the female and male maps formed a sex-average genetic map, which included 4,249 SNPs with a total length of 1956.28 cM and an average genetic distance of 0.46 cM (**Table 2**; **Figure 4**). Among them, LG2 was the longest (228.19 cM), including 648 SNPs, and the average genetic distance was 0.35 cM. Conversely, LG3 was the shortest group (110.56 cM) with 468 SNP tags but a higher resolution of 0.24 cM between adjacent markers on average. There were 120 SNPs in LG9, which was the least in all 11 LGs, with 1.19 cM, the largest average distance between adjacent markers. Of the 11 LGs, the proportions of genetic gaps (≤ 5 cM) ranged from 92.44 to 99.79%, with 96.86% on average. No genetic gaps larger than 10 cM were observed in LG2, LG4 or LG8. The largest gap presented on LG7, which was 23.86 cM (**Table 2**). The detailed information of all the SNPs and the corresponding genetic and physical positions were displayed in **Supplementary Table S4**.

The Spearman correlation coefficient between the genetic map and the reference genome was approximately 0.722 and up to 0.901, indicating that the marker ordering of the genetic map was basically accurate (**Supplementary Table S5**).

TABLE 3 | QTL mapping results.

Year	QTL	LG	Map position		Supporting SNPs	LOD	PVE (%)
			Start (cM)	End (cM)			
2021	qCD8	8	182.843	183.257	3	3.26–3.35	8.80–9.00
2021	qSBD8	8	182.843	183.257	3	3.17–3.31	7.90–8.20
2021	qBSBN7-1	7	119.234	121.26	5	4.29–4.34	10.00–10.10
2021	qBSBN7-2	7	145.122	146.645	3	4.13–4.29	9.60–10.00
2021	qLNS9	9	120.123	131.097	29	3.27–3.65	7.70–8.60
2021	qLNP1	1	92.668	96.534	4	2.88–2.97	6.80–7.00
2021	qLNP2-1	2	19.951	21.234	6	2.65–2.65	6.30–6.30
2021	qLLWS9	9	69.094	85.132	11	3.02–3.36	7.10–7.90
2021	qSLLS3	3	77.053	78.566	5	2.60–2.64	6.20–6.30
2020	qCD8	8	182.843	183.257	3	4.11–4.24	9.20–9.50
2020	qBSBN7-1	7	119.234	148.157	11	3.15–3.58	7.10–8.10
2020	qSI7	7	119.234	146.645	9	2.60–2.83	5.90–6.40
2020	qSI4-1	4	78.288	83.568	16	2.65–2.95	6.00–6.70
2020	qSBD10	10	190.469	202.015	11	3.20–3.41	7.20–7.70
2020	qLNS7	7	85.434	86.947	5	3.05–3.30	6.90–7.50
2020	qLNP2-1	2	19.951	21.234	6	2.72–2.73	6.20–6.20
2020	qLNP2-2	2	56.792	56.792	4	2.56–2.56	5.80–5.80
2020	qLNP2-3	2	61.8	63.34	7	2.51–2.60	5.70–5.90
2020	qLNP2-4	2	67.325	67.325	4	2.55–2.55	5.80–5.80
2020	qLNP2-5	2	82.433	82.433	5	2.50–2.50	5.70–5.70
2020	qLNP2-6	2	100.356	100.94	16	2.50–2.61	5.70–5.90
2020	qLNP2-7	2	123.624	123.624	3	2.55–2.55	5.80–5.80
2020	qLNP2-8	2	146.796	147.817	11	2.50–2.55	5.70–5.80
2020	qLNP2-9	2	197.261	197.261	15	2.55–2.55	5.80–5.80
2020	qLNP7-1	7	81.885	83.9	9	2.52–2.65	5.70–6.00
2020	qLNP7-2	7	95.235	115.701	11	2.75–2.94	6.30–6.70
2020	qLLLS9	9	69.094	69.094	3	5.30–5.30	11.70–11.70
2020	qSLLS11-1	11	211.278	215.045	27	2.50–2.58	5.70–5.90
2020	qSLLS11-2	11	218.617	223.331	11	2.66–2.72	6.10–6.20
2020	qSLLS5	5	70.972	71.265	3	2.67–2.78	6.10–6.30
2020	qSLWS10	10	183.656	185.741	4	3.10–3.22	7.00–7.30
2019	qCD11-1	11	207.47	208.655	13	2.57–2.57	5.80–5.80
2019	qCD11-2	11	211.779	213.514	15	2.55–2.55	5.70–5.70
2019	qCD4-1	4	98.185	110.8	31	2.53–2.91	5.70–6.50
2019	qLLLS4	4	100.7	100.7	5	3.79–3.79	8.40–8.40
2019	qCD4-2	4	116.649	116.649	6	2.51–2.51	5.60–5.60
2019	qCD4-3	4	119.049	119.632	8	2.53–2.57	5.70–5.80
2019	qCD4-4	4	123.824	126.524	15	2.50–2.53	5.60–5.70
2019	qCD4-5	4	137.479	138.232	7	2.50–2.97	5.60–6.60
2019	qCD4-6	4	143.373	144.048	6	2.56–2.65	5.70–6.00
2019	qCD8	8	182.843	184.973	3	2.51–3.00	5.60–6.70
2019	qSI4-2	4	75.123	76.966	12	4.23–4.40	9.30–9.70
2019	qMSH7	7	88.459	89.464	9	2.50–2.77	5.60–6.20
2019	qSBD11	11	36.459	36.459	7	2.54–2.54	5.70–5.70
2019	qLNS8	8	118.779	122.308	4	3.21–3.27	7.20–7.30
2019	qLNP10	10	12.269	12.269	5	3.05–3.05	6.80–6.80
2019	qSLWS10-2	10	10.417	12.269	6	2.72–2.72	6.10–6.10
2019	qLNP2-1	2	19.951	28.282	15	3.06–3.42	6.80–7.60
2019	qLNP2-10	2	57.294	58.434	17	3.05–3.09	6.80–6.90
2019	qLLLS10	10	202.015	202.015	3	3.71–3.71	8.20–8.20
2019	qLLLS11	11	203.027	205.106	7	3.70–3.76	8.20–8.30
2019	qLLWS4	4	115.455	116.649	5	2.64–2.66	5.90–6.00
2019	qSLLS11-3	11	222.728	222.728	3	3.83–3.83	8.50–8.50
2019	qSLWS10-1	10	4.533	5.035	3	2.59–2.59	5.80–5.80

QTLs for Growth and Leaf-Related Traits

We divided the 12 traits into two categories, namely, phenotypes associated with gardenia growth (CD, BSBN, SI, PH, MSH, and SBD) and leaves (LNS, LNP, LLLS, LLWS, SLLS, and SLWS). Using the high-density genetic map and continuous phenotypic data, 23, 22, and 9 QTLs were mapped in 2019, 2020 and 2021, respectively, of which 18 QTLs were associated with the growth traits, while 31 QTLs were related to the leaf-related traits (**Table 3**). These QTLs were distributed in all the LGs of gardenia except LG6, with phenotypic variance explained (PVE) values ranging from 5.6 to 11.7%.

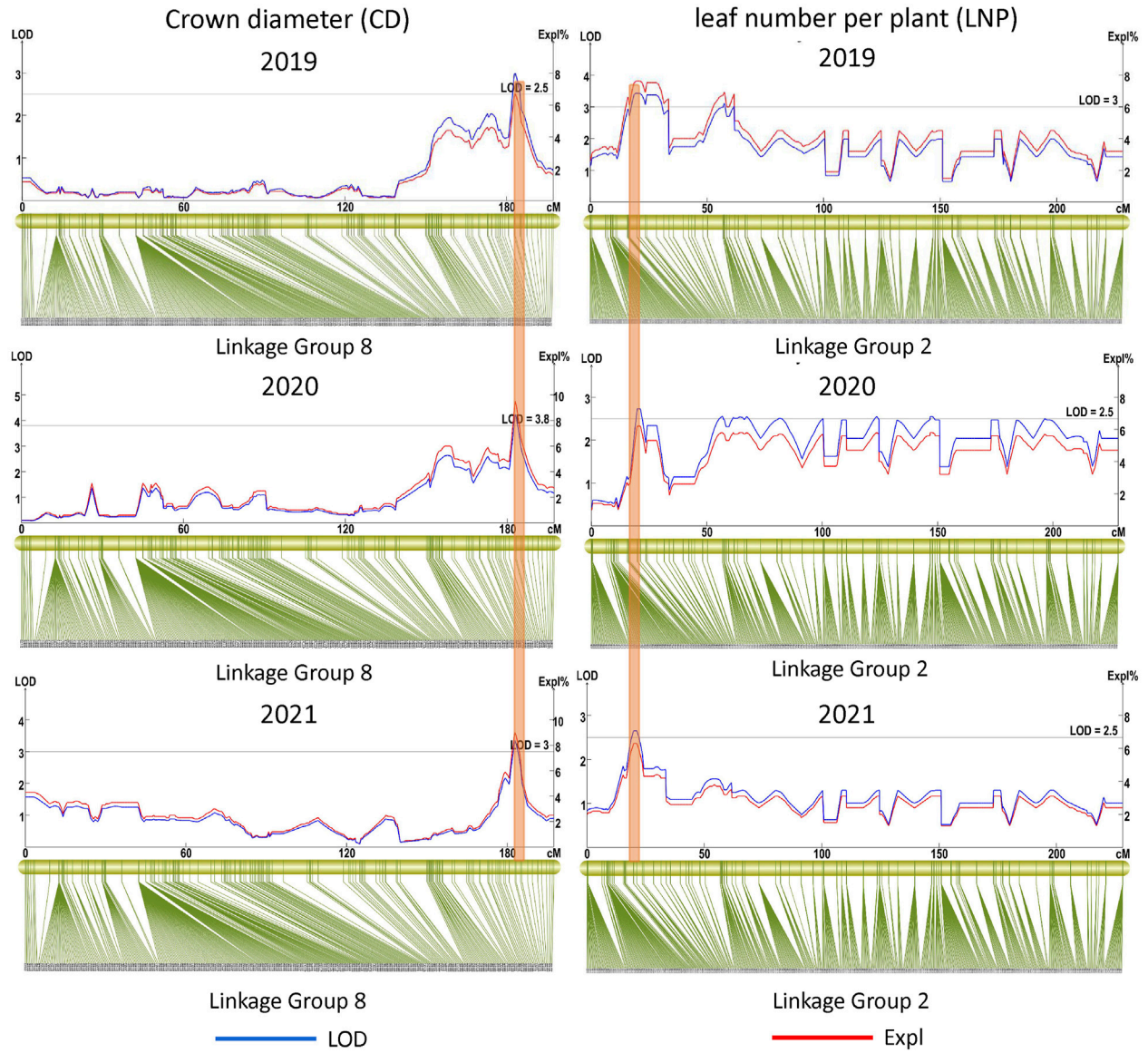


FIGURE 5 | Repeatable QTLs for three dynamic phenotyping time points.

Eighteen QTLs were detected for the six growth traits except PH, including nine for CD, two for BSBN, one for MSH, three for SI, and three for SBD (**Table 3**). Two major-effect QTLs, *qBSBN7-1* and *qBSBN7-2* (PVE = 10–10.1%), were identified with LOD>4, and *qBSBN7-1* could be detected in both 2020 and 2021 (**Table 3**). The LOD of *qCD4-1* in 2019 was slightly over 2.5 but was supported by 31 SNPs. Notably, *qCD8*, which was located in the 182.843–184.973 cM interval, could be repeatably mapped over 3 years, indicating that this QTL could be expressed continuously, contributing to the establishment of the crown (**Figure 5**). Especially in 2020, the LOD value of *qCD8* exceeded four, and PVE exceeded 9%. For SI, there were two QTLs (*qSI4-1* and *qSI4-2*) with a spacing of only 1.322 cM, which were supported by 16 and 12 SNPs, respectively.

In the QTLs of the leaf-related phenotype, four QTLs for LLLS were mapped to LG4, LG9, LG10 and LG11, of which the highest LOD value of *qLLLS9* in 2020 was up to 5.3, and the corresponding PVE was equal to 11.7%. The largest number of identified QTLs belonged to trait LNP, up to 12, and was primarily distributed on LG10, of which *qLNP2-1*, located at 19.951–28.282 cM, was detected for three consecutive years (**Table 3**; **Figure 5**). *qLNP2-2* and *qLNP2-10* were 0.502 cM apart, implying that they might be the same QTL. A total of five QTLs were responsible for SLLS, of which *qSLLS11-1*, *qSLLS11-2* and *qSLLS11-3* gathered between 211.278 and 223.331 cM, with 41 supported SNPs. There were three QTLs for LNS on LG7, LG8 and LG9. With respect to SLWS and LLWS, three and two QTLs were mapped, respectively.

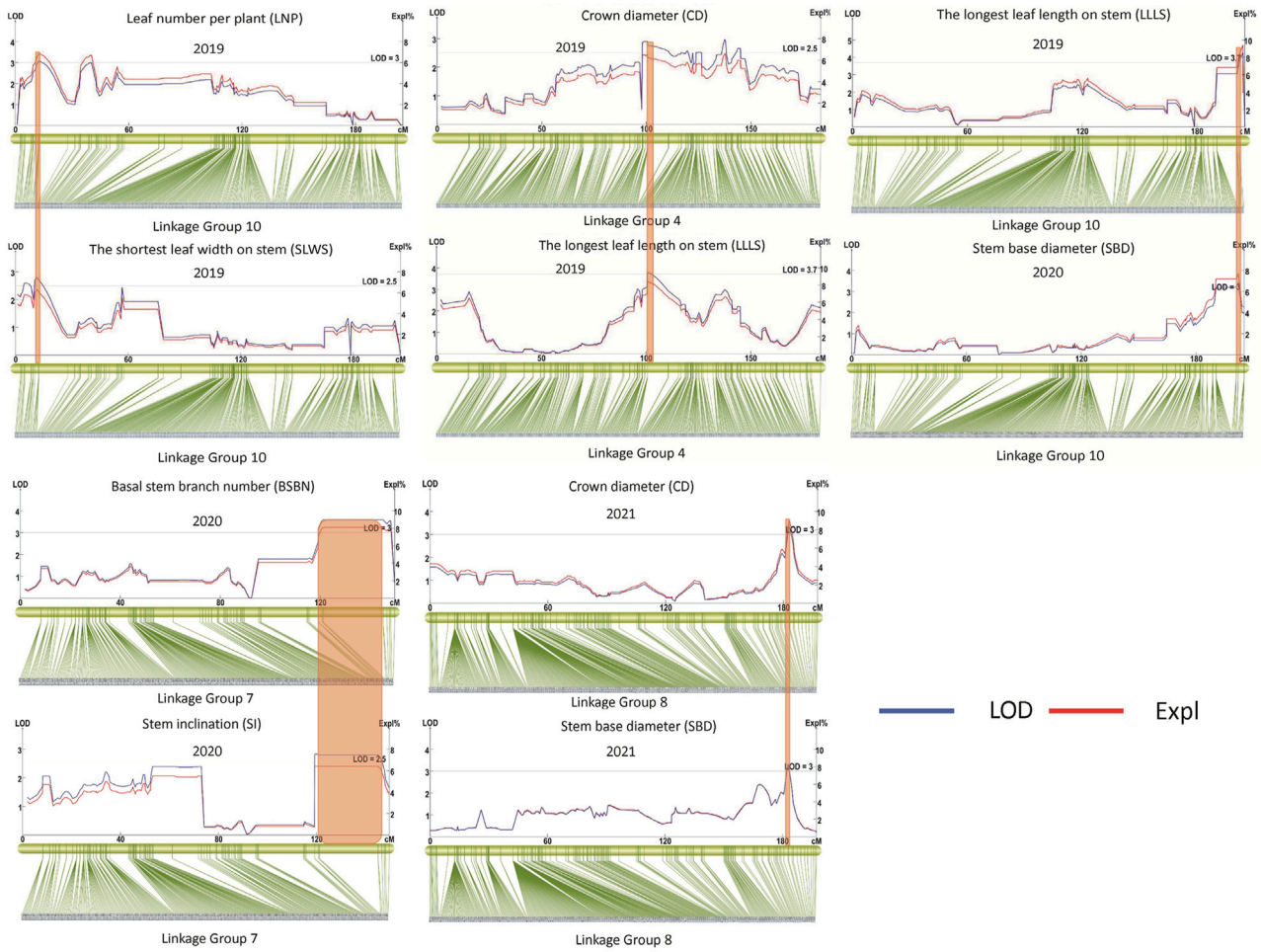


FIGURE 6 | Potential pleiotropism QTLs.

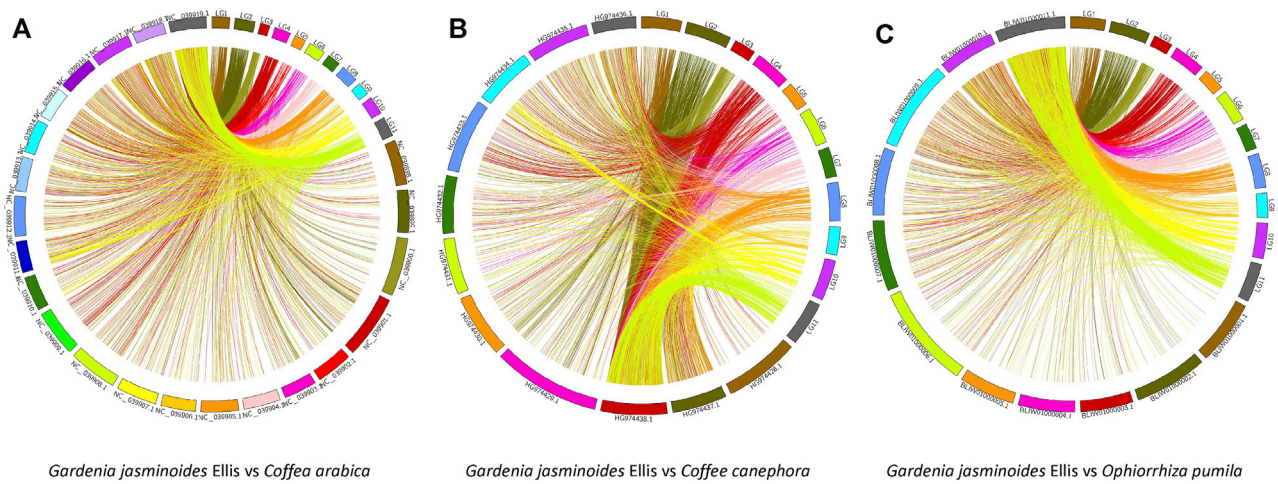


FIGURE 7 | Synteny analyses between the genetic map of gardenia and the genomes of *Coffea arabica* **(A)**, *Coffea canephora* **(B)**, and *Ophiorrhiza pumila* **(C)**.

We continued to explore the QTLs among different phenotypes and found that there were five pairs of QTLs with shared regions, including *qCD8* and *qSBD8*, *qBSBN7* and *qSI7*, *qCD4-1* and *qLLLS4*, *qLNP10* and *qSLWS10-2*, *qSBD10* and *qLLLS10* (**Table 3**; **Figure 6**), suggesting that each pair underlying a single QTL and pleiotropism might play a significant role in gardenia morphogenesis and vegetative development. The structural and functional gene annotations of the above stable and potential pleiotropism QTLs were isolated, resulting in 2,514 nonredundant genes and the corresponding annotation information (**Supplementary Table S6–S7**).

KASP-Based SNP Confirmation

We selected 17 SNP-based KASP markers on chromosomes 2, 7, 9, and 11 and a total of 96 samples for SNP accuracy verification. There were 15 out of 17 markers with successful fluorescence signals in the HC KASP platform, accounting for 88.24%. Among these 15 markers, 13 SNP markers showed genotypes consistent with the GBS results of each individual (**Supplementary Table S8**), indicating the accuracy of the sequencing analysis.

Synteny Analyses

We used this high-density genetic map to investigate the evolutionary relationship of Rubiaceae species, as shown in **Figure 7**. Different levels of synteny were observed between the LGs of gardenia between *C. arabica* (A), *C. canephora* (B) and *O. pumila* (C). Specifically, relatively strong synteny was consistently noted between LG11 of gardenia and NC_039,919.1 of *C. arabica*, HG974438.1 of *C. canephora*, and Bliw01000011.1 of *O. pumila*, indicating that LG11 was more conserved than other LGs. In addition, this type of stronger collinearity was also found between these pairs: the pair LG10 and *C. arabica*'s chromosome NC_039,919.1, the pair LG4 and *C. canephora*'s chromosome HG974438.1, and the pair LG10 and *O. pumila*'s chromosome BLIW01000011.1. Moreover, chromosomes NC_039,917.1, NC_039,918.1 and NC_039,919.1 demonstrated higher collinearity than other chromosomes in *C. arabica*. Similarly, chromosomes HG974438.1, HG974437.1 and *HG974436.1* in *C. canephora* and chromosomes BLIW01000010.1 and BLIW01000011.1 in *O. pumila* displayed stronger synteny than other corresponding chromosomes.

DISCUSSION

Gardenia is a type of gardening species that has medicinal and industrial value. At present, there is less genetic research on this species. In this study, after constructing an F_1 segregating population, the first high-density genetic map of gardenia was accomplished using a high-throughput sequencing method.

Three-year dynamic QTL positioning identified a panel of vegetative growth-related QTLs. We believe that this research will open a new avenue for gardenia molecular genetic research.

NGS has greatly accelerated the process of QTL mapping according to forward genetics, as primarily performed by bulk segregation analysis (BSA), high-density genetic map-based QTL mapping and genome-wide association studies (GWASs) in horticultural plants (Ban and Xu, 2020; Ferrão et al., 2020; Song et al., 2020). GBS, which is rooted in NGS, has opened up new possibilities for genome-wide SNP mining without high investments (Chung et al., 2017). At the cost of approximately 1.8 Gb sequencing data per sample, 154,909 dual-calling SNPs were *de novo* developed in the present study. Hereafter, the sequencing depth, segregation distortion and integrity control were processed to guarantee a high-quality panel of SNPs for high-density genetic map construction. Using an F_1 population of 200 plants of gardenia and GBS-based genotyping, a high-density genetic map harboring 4,249 SNPs was constructed, which showed high resolution (0.46 cM per adjacent SNPs) and satisfied marker orders with an approximately 0.8 collinearity compared with the reference genome, a similar standard as in other species (Ji et al., 2018; Lu et al., 2019; Wang et al., 2020). To evaluate the SNP accuracy, we selected 17 SNPs and genotyped 96 F_1 individuals, and high consistency between the KASP and GBS genotypes was observed according to Song's report (Song et al., 2020). These results indicated that a high-quality and high-density genetic map was generated. Furthermore, the genetic map was used for the QTL mapping responsible for early growth and development traits for three dynamic years, and a total of 49 QTLs for 12 traits (CD, BSBN, SI, PH, MSH, SBD, LNS, LNP, LLLS, LLWS, SLLS, and SLWS) were identified. Because the most useful organ was the gardenia fruit, further QTLs associated with fruit-related traits, such as the weight, shape, size or functional substance content, could be expected in 2022 or later, when all F_1 gardenia individuals will transfer to reproductive growth. This high-density genetic map might provide a new lesson for molecular genetic research in gardenia.

As mentioned above, 49 QTLs for 12 traits were mapped, including three stably expressed QTLs, *qBSBN7-1*, *qCD8* and *qLNP2-1* (**Table 3**). These QTLs played persistent roles in the corresponding morphogenesis during the vegetative period and were valuable and useful for MAS-based breeding programs (Jamshed et al., 2016; Galiano-Carneiro et al., 2021). In addition, we also found five regions within two QTLs (*qCD8* and *qSBD8*, *qBSBN7* and *qSI7*, *qCD4-1* and *qLLLS4*, *qLNP10* and *qSLWS10-2*, *qSBD10* and *qLLLS10*), suggesting that each region underlies a single QTL with pleiotropism (Sun et al., 2017; Wang et al., 2018). Notably, there was a genetic basis for the phenotypic correlation between CD and SBD, and CD and LLLS, which was consistent with the strong correlations between CD and SBD (0.40***), and CD and LLLS (0.53***) (**Figure 2**). The SNPs underlying these stable and pleiotropic QTLs could be further converted into KASP markers and potentially used as MAS markers. Further gene cloning may also benefit from the gene structural and functional annotations underlying stable and potential pleiotropism QTLs (**Supplementary Table S6, S7**).

Generally, gardenia plants are focused on vegetative development for the first 3 years after seed germination, which is called the juvenile period. A gradual declining trend in QTL numbers was observed from 2019, 2020 to 2021, which might be associated with the fact that the channels of vegetative development gradually slowed, while reproductive growth gradually opened. Furthermore, some year-specific QTLs were found except stable and pleiotropism potential QTLs, which could be explained by specific functions in plant growth phases. This phenotyping-derived dynamic QTL mapping based on continuous development time points has recently been performed in

peaches (Desnoues et al., 2016), *Populus* (Du et al., 2019), *Catalpa bungei* (Lu et al., 2019), and chrysanthemum (Ao et al., 2019). Compared with these studies, one shortcoming of this study is that phenotyping was conducted at only three time points, which might limit the dissection resolution of the developmental trait inheritance. Currently, high-throughput phenomics combining spectral imaging and machine learning methods provides particular insight into the deciphering of dynamic phenotypes in a way that is plant damage-free (Li and Sillanpää 2015; Adams et al., 2020; Streich et al., 2020; Zhu et al., 2021). Novel dynamic QTL perspectives might be enabled by employing phenomic methods in the near future.

The complex synteny of gardenia with *C. arabica*, *C. canephora* and *O. pumila* (**Figure 7**) implied that widely chromosomal fission and fusion have happened after their divergence from the common ancestor, similar to other species (Luo et al., 2020; Yang et al., 2020). The synteny levels of all LGs in gardenia indicated that different chromosomes underwent different evolutionary process. Strong synteny in LG10, LG11 and LG4 of gardenia demonstrated that these LGs were more conserved than other LGs. The sequences underlying these conserved regions could be potentially used to speculate the corresponding genetic information of species in *Salicaceae*, and the QTLs on LG4, LG10 and LG11 (*qCD4-1*, *qSBD10*, *qSLLS11-3*, etc.), potentially orthologous QTLs (Rinaldi et al., 2016; Webb et al., 2016), could be applied for comparative mapping in other species of *Rubiaceae*. Ta˚Cken together, the synteny analyses of this paper may lay a foundation for subsequent comparative genomic research.

One ultimate goal of QTL mapping is to perform QTL fine mapping, screening and gene cloning of candidate genes. This goal was commonly achieved in therophyte plants, such as through map-based cloning (Jia et al., 2020; Yu et al., 2020; Sierra-Orozco et al., 2021). However, for many perennial species, it takes more than 6 years to perform hybridization and further backcrossing, and the population size is always restricted to several hundred, which leads to limited recombination. This limitation makes the process of QTL cloning in perennial species slow. To expedite this process, transcriptomics analysis can be used to call RNA variants and differentially expressed genes (DEGs) within QTL regions (Park et al., 2019; Wen et al., 2019). Other omics analyses, such as metabolomics and proteomics, can also provide useful information on the metabolic chemicals or proteins related to phenotypic variations (Szymański et al., 2020; Mou et al., 2021). In future experimental designs, these analyses will be considered to understand the basis of phenotypic variation comprehensively.

5 CONCLUSION

In this study, we developed a panel of genome-wide high-quality SNPs using the GBS method and providing the first high-density genetic map in the gardenia. SNPs and genetic maps could be useful for further genetic study and evolutionary genomics. Based on this high-density genetic map, 18 and 31 QTLs were identified for growth traits and leaf-related traits at three dynamic phenotyping time points, respectively. Stably expressed QTLs and potential pleiotropism QTLs could be targets for MAS breeding and for further gene cloning.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

XW conceived and designed the experiments; YC performed paternity test, phenotying and first draft writing; BF and SS performed phenotying; YX performed the data analysis; YC wrote the manuscript and XW, YX revised the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant number 31660218) and the Scientific Foundation of Double First-class Discipline Development of TCM (grant number JXSYLXK-ZHYA0028). National Natural Science Foundation of China is a national public funder. Scientific Foundation of Double First-class Discipline Development of TCM is a funder of Jiangxi University of Chinese Medicine.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.802738/full#supplementary-material

REFERENCES

Adams, J., Qiu, Y., Xu, Y., and Schnable, J. C. (2020). Plant Segmentation by Supervised Machine Learning Methods. *Plant phenome j.* 3, e20001. doi:10.1002/ppj2.20001

Anagbogu, C. F., Bhattacharjee, R., Ilori, C., Tongyoo, P., Dada, K. E., Muyiwa, A. A., et al. (2019). Genetic Diversity and Re-classification of Coffee (*Coffea Canephora* Pierre Ex A. Froehner) from South Western Nigeria through Genotyping-By-Sequencing-Single Nucleotide Polymorphism Analysis. *Genet. Resour. Crop Evol.* 66, 685–696. doi:10.1007/s10722-019-00744-2

Ao, N., Ma, J., Xu, T., Su, J., Yang, X., Guan, Z., et al. (2019). Genetic Variation and QTL Mapping for Cold Tolerance in a chrysanthemum F1 Population at Different Growth Stages. *Euphytica* 215, 88. doi:10.1007/s10681-019-2412-7

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., et al. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One* 3, e3376. doi:10.1371/journal.pone.0003376

Ban, S., and Xu, K. (2020). Identification of Two QTLs Associated with High Fruit Acidity in Apple Using Pooled Genome Sequencing Analysis. *Hortic. Res.* 7, 171. doi:10.1038/s41438-020-00393-y

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., et al. (2004). The Pfam Protein Families Database. *Nucleic Acids Res.* 32, 138D–141D. doi:10.1093/nar/gkh121

Berlin, S., Lagercrantz, U., von Arnold, S., Öst, T., and Rönnberg-Wästljung, A. (2010). High-density Linkage Mapping and Evolution of Paralogs and Orthologs in Salix and Populus. *BMC Genomics* 11, 129. doi:10.1186/1471-2164-11-129

Chang, Y., Ding, J., Xu, Y., Li, D., Zhang, W., Li, L., et al. (2018). SLAF-based High-Density Genetic Map Construction and QTL Mapping for Major Economic Traits in Sea Urchin *Strongylocentrotus Intermedius*. *Sci. Rep.* 8, 820. doi:10.1038/s41598-017-18768-y

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020c). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13, 1194–1202. doi:10.1016/j.molp.2020.06.009

Chen, L., Li, M., Yang, Z., Tao, W., Wang, P., Tian, X., et al. (2020b). *Gardenia Jasminoides* Ellis: Ethnopharmacology, Phytochemistry, and Pharmacological and Industrial Applications of an Important Traditional Chinese Medicine. *J. Ethnopharmacology* 257, 112829. doi:10.1016/j.jep.2020.112829

Chen, Q., Xue, G., Ni, Q., Wang, Y., Gao, Q., Zhang, Y., et al. (2020a). Physicochemical and Rheological Characterization of Pectin-rich Polysaccharides from Gardenia Jasminoides J. Ellis Flower. *Food Sci. Nutr.* 8, 3335–3345. doi:10.1002/fsn3.1612

Chung, Y. S., Choi, S. C., Jun, T.-H., and Kim, C. (2017). Genotyping-by-Sequencing: a Promising Tool for Plant Genetics Research and Breeding. *Hortic. Environ. Biotechnol.* 58, 425–431. doi:10.1007/s13580-017-0297-8

Deng, S.-Y., Wang, X.-r., Zhu, P.-l., Wen, Q., and Yang, C.-x. (2015). Development of Polymorphic Microsatellite Markers in the Medicinal Plant *Gardenia Jasminoides* (*Rubiaceae*). *Biochem. Syst. Ecol.* 58, 149–155. doi:10.1016/j.bse.2014.11.009

Desnoues, E., Baldazzi, V., Génard, M., Mauroux, J.-B., Lambert, P., Confolent, C., et al. (2016). Dynamic QTLs for Sugars and Enzyme Activities Provide an Overview of Genetic Control of Sugar Metabolism during Peach Fruit Development. *Exbotj* 67, 3419–3431. doi:10.1093/jxb/erw169

Dong, M., He, Q., Zhao, J., Zhang, Y., Yuan, D., and Zhang, A. J. (2019). Genetic Mapping of Prince Rupprecht's Larch (*Larix Principis-rupprechtii* Mayr) by Specific-Locus Amplified Fragment Sequencing. *Genes* 10, 583. doi:10.3390/genes10080583

Du, Q., Yang, X., Xie, J., Quan, M., Xiao, L., Lu, W., et al. (2019). Time-specific and Pleiotropic Quantitative Trait Loci Coordinately Modulate Stem Growth inPopulus. *Plant Biotechnol. J.* 17, 608–624. doi:10.1111/pbi.13002

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

Ferrão, L. F. V., Johnson, T. S., Benevenuto, J., Edger, P. P., Colquhoun, T. A., and Munoz, P. R. (2020). Genome-wide Association of Volatiles Reveals Candidate Loci for Blueberry Flavor. *New Phytol.* 226, 1725–1737. doi:10.1111/nph.16459

Gabay, G., Dahan, Y., Izhaki, Y., Faigenboim, A., Ben-Ari, G., Elkind, Y., et al. (2018). High-resolution Genetic Linkage Map of European Pear (*Pyrus Communis*) and QTL fine-mapping of Vegetative Budbreak Time. *BMC Plant Biol.* 18, 175. doi:10.1186/s12870-018-1386-2

Galiano-Carneiro, A. L., Kessel, B., Presterl, T., and Miedaner, T. (2021). Intercontinental Trials Reveal Stable QTL for Northern Corn Leaf Blight Resistance in Europe and in Brazil. *Theor. Appl. Genet.* 134, 63–79. doi:10.1007/s00122-020-03682-1

Hanley, S. J., Mallott, M. D., and Karp, A. (2006). Alignment of a *Salix* Linkage Map to the *Populus* Genomic Sequence Reveals Macrosynteny between Willow and poplar Genomes. *Tree Genet. Genomes* 3, 35–48. doi:10.1007/s11295-006-0049-x

Higashino, S., Sasaki, Y., Giddings, J. C., Hyodo, K., Fujimoto Sakata, S., Matsuda, K., et al. (2014). Crocetin, a Carotenoid fromGardenia jasminoidesEllis, Protects against Hypertension and Cerebral Thrombogenesis in Stroke-Prone Spontaneously Hypertensive Rats. *Phytother. Res.* 28, 1315–1319. doi:10.1002/ptr.5130

Hu, Y., Liu, X., Xia, Q., Yin, T., Bai, C., Wang, Z., et al. (2019). Comparative Anti-arthritic Investigation of Iridoid Glycosides and Crocetin Derivatives from *Gardenia Jasminoides* Ellis in Freund's Complete Adjuvant-Induced Arthritis in Rats. *Phytomedicine* 53, 223–233. doi:10.1016/j.phymed.2018.07.005

İpek, A., İpek, M., Ercişli, S., and Tangu, N. A. (2017). Transcriptome-based SNP Discovery by GBS and the Construction of a Genetic Map for Olive. *Funct. Integr. Genomics* 17, 493–501. doi:10.1007/s10142-017-0552-1

İpek, A., Yılmaz, K., Sıkıcı, P., Tangu, N. A., Öz, A. T., Bayraktar, M., et al. (2016). SNP Discovery by GBS in Olive and the Construction of a High-Density Genetic Linkage Map. *Biochem. Genet.* 54, 313–325. doi:10.1007/s10528-016-9721-5

Jamshed, M., Jia, F., Gong, J., Palanga, K. K., Shi, Y., Li, J., et al. (2016). Identification of Stable Quantitative Trait Loci (QTLs) for Fiber Quality Traits across Multiple Environments in *Gossypium Hirsutum* Recombinant Inbred Line Population. *BMC Genomics* 17, 197. doi:10.1186/s12864-016-2560-2

Ji, F., Wei, W., Liu, Y., Wang, G., Zhang, Q., Xing, Y., et al. (2018). Construction of a SNP-Based High-Density Genetic Map Using Genotyping by Sequencing (GBS) and QTL Analysis of Nut Traits in Chinese Chestnut (*Castanea Mollissima* Blume). *Front. Plant Sci.* 9, 816. doi:10.3389/fpls.2018.00816

Jia, H., Li, M., Li, W., Liu, L., Jian, Y., Yang, Z., et al. (2020). A Serine/threonine Protein Kinase Encoding Gene *KERNEL NUMBER PER ROW6* Regulates maize Grain Yield. *Nat. Commun.* 11, 988. doi:10.1038/s41467-020-14746-7

Jorge, V., Dowkiw, A., Faivre-Rampant, P., and Bastien, C. (2005). Genetic Architecture of Qualitative and Quantitative *Melampsora Larici-populina* Leaf Rust Resistance in Hybrid poplar: Genetic Mapping and QTL Detection. *New Phytol.* 167, 113–127. doi:10.1111/j.1469-8137.2005.01424.x

Kai, W., Kikuchi, K., Tohari, S., Chew, A. K., Tay, A., Fujiwara, A., et al. (2011). Integration of the Genetic Map and Genome Assembly of Fugu Facilitates Insights into Distinct Features of Genome Evolution in Teleosts and Mammals. *Genome Biol. Evol.* 3, 424–442. doi:10.1093/gbe/evr041

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Kent, W. J. (2002). BLAT-the BLAST-like Alignment Tool. *Genome Res.* 12, 656–664. doi:10.1101/gr.229202

Khajeh, E., Rasmi, Y., Kheradmand, F., Malekinejad, H., Aramwit, P., Saboory, E., et al. (2020). Crocetin Suppresses the Growth and Migration in HCT-116 Human Colorectal Cancer Cells by Activating the P-38 MAPK Signaling Pathway. *Res. Pharma Sci.* 15, 592–601. doi:10.4103/1735-5362.301344

Kim, B., Hwang, I. S., Lee, H. J., Lee, J. M., Seo, E., Choi, D., et al. (2018). Identification of a Molecular Marker Tightly Linked to Bacterial Wilt Resistance in Tomato by Genome-wide SNP Analysis. *Theor. Appl. Genet.* 131, 1017–1030. doi:10.1007/s00122-018-3054-1

Kodama, M., Brieuc, M. S. O., Devlin, R. H., Hard, J. J., and Naish, K. A. (2014). Comparative Mapping between Coho salmon (*Oncorhynchus kisutch*) and Three Other Salmonids Suggests a Role for Chromosomal Rearrangements in the Retention of Duplicated Regions Following a Whole Genome Duplication Event. *G3-genes Genom. Genet.* 4, 1717–1730. doi:10.1534/g3.114.012294

Kosambi, D. D. (1943). The Estimation of Map Distances from Recombination Values. *Ann. Eugen.* 12, 172–175. doi:10.1111/j.1469-1809.1943.tb02321.x

Lambert, P., Dicenta, F., Rubio, M., and Audergon, J. M. (2007). QTL Analysis of Resistance to Sharka Disease in the Apricot (Prunus Armeniaca L.) 'Polonais' × 'Stark Early Orange' F1 Progeny. *Tree Genet. Genomes* 3, 299–309. doi:10.1007/s11295-006-0069-6

Lewter, J., Worthington, M. L., Clark, J. R., Varanasi, A. V., Nelson, L., Owens, C. L., et al. (2019). High-density Linkage Maps and Loci for berry Color and Flower Sex in Muscadine Grape (*Vitis Rotundifolia*). *Theor. Appl. Genet.* 132, 1571–1585. doi:10.1007/s00122-019-03302-7

Li, H., and Durbin, R. (2009). Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

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

Li, W., Ren, C., Fei, C., Wang, Y., Xue, Q., Li, L., et al. (2021). Analysis of the Chemical Composition Changes of Gardeniae Fructus before and after Processing Based on Ultra-high-performance Liquid Chromatography Quadrupole Time-of-flight Mass Spectrometry. *J. Sep. Sci.* 44, 981–991. doi:10.1002/jssc.202000957

Li, Z., and Sillanpää, M. J. (2015). Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends Plant Sci.* 20, 822–833. doi:10.1016/j.tplants.2015.08.012

Lu, N., Zhang, M., Xiao, Y., Han, D., Liu, Y., Zhang, Y., et al. (2019). Construction of a High-Density Genetic Map and QTL Mapping of Leaf Traits and Plant Growth in an Interspecific F1 Population of Catalpa Bungei × Catalpa Duclouxii Dode. *BMC Plant Biol.* 19, 596. doi:10.1186/s12870-019-2207-y

Luo, X., Xu, L., Wang, Y., Dong, J., Chen, Y., Tang, M., et al. (2020). An Ultra-high-density Genetic Map Provides Insights into Genome Synteny, Recombination Landscape and Taproot Skin Colour in Radish (Raphanus Sativus L). *Plant Biotechnol. J.* 18, 274–286. doi:10.1111/pbi.13195

Mathew, L. S., Spannagl, M., Al-Malki, A., George, B., Torres, M. F., Al-Dous, E. K., et al. (2014). A First Genetic Map of Date palm (*Phoenix Dactylifera*) Reveals Long-Range Genome Structure Conservation in the Palms. *BMC Genomics* 15, 285. doi:10.1186/1471-2164-15-285

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Mohler, V., and Stadlmeier, M. (2019). Dynamic QTL for Adult Plant Resistance to Powdery Mildew in Common Wheat (*Triticum aestivum* L). *J. Appl. Genet.* 60, 291–300. doi:10.1007/s13353-019-00518-7

Moncada, M. D. P., Tovar, E., Montoya, J. C., González, A., Spindel, J., and McCouch, S. (2016). A Genetic Linkage Map of Coffee (*Coffea Arabica* L.) and QTL for Yield, Plant Height, and Bean Size. *Tree Genet. Genomes* 12, 5. doi:10.1007/s11295-015-0927-1

Mou, J., Zhang, Z., Qiu, H., Lu, Y., Zhu, X., Fan, Z., et al. (2021). Multiomics-based Dissection of Citrus Flavonoid Metabolism Using a Citrus Reticulata × Poncirus Trifoliata Population. *Hortic. Res.* 8, 56. doi:10.1038/s41438-021-00472-8

Oyant, L. H.-S., Crespel, L., Rajapakse, S., Zhang, L., and Foucher, F. (2007). Genetic Linkage Maps of Rose Constructed with New Microsatellite Markers and Locating QTL Controlling Flowering Traits. *Tree Genet. Genomes* 4, 11–23. doi:10.1007/s11295-007-0084-2

Pacheco, I., Bassi, D., Eduardo, I., Ciacciulli, A., Pirona, R., Rossini, L., et al. (2014). QTL Mapping for Brown Rot (Monilinia Fructigena) Resistance in an Intraspecific Peach (Prunus Persica L. Batsch) F1 Progeny. *Tree Genet. Genomes* 10, 1223–1242. doi:10.1007/s11295-014-0756-7

Park, M., Lee, J.-H., Han, K., Jang, S., Han, J., Lim, J.-H., et al. (2019). A Major QTL and Candidate Genes for Capsaicinoid Biosynthesis in the Pericarp of *Capsicum Chinense* Revealed Using QTL-Seq and RNA-Seq. *Theor. Appl. Genet.* 132, 515–529. doi:10.1007/s00122-018-3238-8

Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient Polyploidization Predating Divergence of the Cereals, and its Consequences for Comparative Genomics. *Proc. Natl. Acad. Sci.* 101, 9903–9908. doi:10.1073/pnas.0307901101

Paudel, D., Kannan, B., Yang, X., Harris-Shultz, K., Thudi, M., Varshney, R. K., et al. (2018). Surveying the Genome and Constructing a High-Density Genetic Map of Napiergrass (*Cenchrus Purpureus* Schumach). *Sci. Rep.* 8, 14419. doi:10.1038/s41598-018-32674-x

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for *De Novo* SNP Discovery and Genotyping in Model and Non-model Species. *PLoS One* 7, e37135. doi:10.1371/journal.pone.0037135

Pootakham, W., Ruang-Areerate, P., Jomchai, N., Sonthirod, C., Sangsrakru, D., Yoocha, T., et al. (2015). Construction of a High-Density Integrated Genetic Linkage Map of Rubber Tree (*Hevea Brasiliensis*) Using Genotyping-By-Sequencing (GBS). *Front. Plant Sci.* 6, 367. doi:10.3389/fpls.2015.00367

Qin, F.-m., Meng, L.-j., Zou, H.-l., and Zhou, G.-x. (2013). Three New Iridoid Glycosides from the Fruit of Gardenia Jasminoides Var. Radicans. *Chem. Pharm. Bull.* 61, 1071–1074. doi:10.1248/cpb.c13-00262

Rehman, F., Gong, H., Li, Z., Zeng, S., Yang, T., Ai, P., et al. (2020). Identification of Fruit Size Associated Quantitative Trait Loci Featuring SLAF Based High-Density Linkage Map of Goji berry (*Lycium* spp). *BMC Plant Biol.* 20, 474. doi:10.1186/s12870-020-02567-1

Rinaldi, R., Van Deynze, A., Portis, E., Rotino, G. L., Toppino, L., Hill, T., et al. (2016). New Insights on Eggplant/tomato/pepper Synteny and Identification of Eggplant and Pepper Orthologous QTL. *Front. Plant Sci.* 7, 1031. doi:10.3389/fpls.2016.01031

Robledo, D., Palaiokostas, C., Bargelloni, L., Martínez, P., and Houston, R. (2018). Applications of Genotyping by Sequencing in Aquaculture Breeding and Genetics. *Rev. Aquacult.* 10, 670–682. doi:10.1111/raq.12193

Rubio, B., Lalanne-Tisné, G., Voisin, R., Tandonnet, J.-P., Portier, U., Van Ghelder, C., et al. (2020). Characterization of Genetic Determinants of the Resistance to Phylloxera, *Daktulosphaira Vitifoliae*, and the Dagger Nematode *Xiphinema index* from Muscadine Background. *BMC Plant Biol.* 20, 213. doi:10.1186/s12870-020-2310-0

Sánchez-Pérez, R., Dicenta, F., and Martínez-Gómez, P. (2012). Inheritance of Chilling and Heat Requirements for Flowering in almond and QTL Analysis. *Tree Genet. Genomes* 8, 379–389. doi:10.1007/s11295-011-0448-5

Schneider, M., Tognolli, M., and Bairoch, A. (2004). The Swiss-Prot Protein Knowledgebase and ExPASy: Providing the Plant Community with High Quality Proteomic Data and Tools. *Plant Physiol. Biochem.* 42, 1013–1021. doi:10.1016/j.plaphy.2004.10.009

Shang, L., Liu, F., Wang, Y., Abduweli, A., Cai, S., Wang, K., et al. (2015). Dynamic QTL Mapping for Plant Height in Upland Cotton (*Gossypium Hirsutum*). *Plant Breed* 134, 703–712. doi:10.1111/pbr.12316

Sierra-Orozco, E., Shekasteband, R., Illa-Berenguer, E., Snouffer, A., van der Knaap, E., Lee, T. G., et al. (2021). Identification and Characterization of *GLOBE*, a Major Gene Controlling Fruit Shape and Impacting Fruit Size and Marketability in Tomato. *Hortic. Res.* 8, 138. doi:10.1038/s41438-021-00574-3

Song, X., Xu, Y., Gao, K., Fan, G., Zhang, F., Deng, C., et al. (2020). High-density Genetic Map Construction and Identification of Loci Controlling Flower-type Traits in Chrysanthemum (Chrysanthemum × Morifolium Ramat). *Hortic. Res.* 7, 108. doi:10.1038/s41438-020-0333-1

Streich, J., Romero, J., Gazolla, J. G. F. M., Kainer, D., Cliff, A., Prates, E. T., et al. (2020). Can Exascale Computing and Explainable Artificial Intelligence Applied to Plant Biology Deliver on the United Nations Sustainable Development Goals? *Curr. Opin. Biotechnol.* 61, 217–225. doi:10.1016/j.copbio.2020.01.010

Sun, B., Zhan, X.-D., Lin, Z.-C., Wu, W.-X., Yu, P., Zhang, Y.-X., et al. (2017). Fine Mapping and Candidate Gene Analysis of *qHD5*, a Novel Major QTL with Pleiotropism for Yield-Related Traits in rice (*Oryza Sativa* L). *Theor. Appl. Genet.* 130, 247–258. doi:10.1007/s00122-016-2787-y

Sun, J., Wang, J., Liu, H., Xie, D., Zheng, H., Zhao, H., et al. (2015). Dynamic QTL Analysis of rice Seedling Height and Tiller Number under Salt Stress. *J. Nucl. Agric. Sci.* 29, 235–243. (In Chinese). doi:10.11869/j.issn.100-8551.2015.02.0235

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., et al. (2013). SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. *PLoS One* 8, e58700. doi:10.1371/journal.pone.0058700

Szymański, J., Bocobza, S., Panda, S., Sonawane, P., Cárdenas, P. D., Lashbrooke, J., et al. (2020). Analysis of Wild Tomato Introgression Lines Elucidates the Genetic Basis of Transcriptome and Metabolome Variation Underlying Fruit Traits and Pathogen Response. *Nat. Genet.* 52, 1111–1121. doi:10.1038/s41588-020-0690-6

Tello, J., Roux, C., Chouiki, H., Laucou, V., Sarah, G., Weber, A., et al. (2019). A Novel High-Density grapevine (*Vitis vinifera* L.) Integrated Linkage Map Using GBS in a Half-Diallel Population. *Theor. Appl. Genet.* 132, 2237–2252. doi:10.1007/s00122-019-03351-y

Tian, Y., Pu, X., Yu, H., Ji, A., Gao, R., Hu, Y., et al. (2020). Genome-wide Characterization and Analysis of bHLH Transcription Factors Related to Crocin Biosynthesis in *Gardenia Jasminoides* Ellis (*Rubiaceae*). *Biomed. Res. Int.* 2020, 1–11. doi:10.1155/2020/2903861

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., et al. (2013). ezRAD: a Simplified Method for Genomic Genotyping in Non-model Organisms. *PeerJ* 1, e203. doi:10.7717/peerj.203

Tsanakas, G. F., Polidoros, A. N., and Economou, A. S. (2013). Genetic Variation in Gardenia Grown as Pot Plant in Greece. *Scientia Horticulturae* 162, 213–217. doi:10.1016/j.scienta.2013.08.020

Van Ooijen, J. W. (2006). *JoinMap ® 4, Software for the Calculation of Genetic Linkage Maps in Experimental Populations*. Wageningen, Netherlands: Kyazma BV.

Van Ooijen, J. W. (2009). *MapQTL ® 6, Software for the Mapping of Quantitative Trait Loci in Experimental Populations of Diploid Species*. Wageningen, Netherlands: Kyazma BV.

van Os, H., Stam, P., Visser, R. G. F., and van Eck, H. J. (2005). SMOOTH: a Statistical Method for Successful Removal of Genotyping Errors from High-Density Genetic Linkage Data. *Theor. Appl. Genet.* 112, 187–194. doi:10.1007/s00122-005-0124-y

Wang, C. M., Lo, L. C., Zhu, Z. Y., and Yue, G. H. (2006). A Genome Scan for Quantitative Trait Loci Affecting Growth-Related Traits in an F_1 Family of Asian Seabass (*Lates calcarifer*). *BMC Genomics* 7, 274. doi:10.1186/1471-2164-7-274

Wang, F., Miao, M., Xia, H., Yang, L.-G., Wang, S.-K., and Sun, G.-J. (2017). Antioxidant Activities of Aqueous Extracts from 12 Chinese Edible Flowers *In Vitro* and *In Vivo*. *Food Nutr. Res.* 61, 1265324. doi:10.1080/16546628.2017.1265324

Wang, J., Li, Q., Zhong, X., Song, J., Kong, L., and Yu, H. (2018). An Integrated Genetic Map Based on EST-SNPs and QTL Analysis of Shell Color Traits in Pacific Oyster *Crassostrea gigas*. *Aquaculture* 492, 226–236. doi:10.1016/j.aquaculture.2018.04.018

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. *Nucleic Acids Res.* 38, e164. doi:10.1093/nar/gkq603

Wang, S., Meyer, E., McKay, J. K., and Matz, M. V. (2012). 2b-RAD: a Simple and Flexible Method for Genome-wide Genotyping. *Nat. Methods* 9, 808–810. doi:10.1038/nmeth.2023

Wang, X., Wang, H., Long, Y., Liu, L., Zhao, Y., Tian, J., et al. (2015). Dynamic and Comparative QTL Analysis for Plant Height in Different Developmental Stages of *Brassica Napus* L. *Theor. Appl. Genet.* 128, 1175–1192. doi:10.1007/s00122-015-2498-9

Wang, X., Zhang, R., Song, W., Han, L., Liu, X., Sun, X., et al. (2019). Dynamic Plant Height QTL Revealed in maize through Remote Sensing Phenotyping Using a High-Throughput Unmanned Aerial Vehicle (UAV). *Sci. Rep.* 9, 3458. doi:10.1038/s41598-019-39448-z

Wang, Y., Wang, C., Han, H., Luo, Y., Wang, Z., Yan, C., et al. (2020). Construction of a High-Density Genetic Map and Analysis of Seed-Related Traits Using Specific Length Amplified Fragment Sequencing for *Cucurbita Maxima*. *Front. Plant Sci.* 10, 1782. doi:10.3389/fpls.2019.01782

Webb, A., Cottage, A., Wood, T., Khamassi, K., Hobbs, D., Gostkiewicz, K., et al. (2016). A SNP-Based Consensus Genetic Map for Synteny-Based Trait Targeting in Faba Bean (*Vicia faba* L). *Plant Biotechnol. J.* 14, 177–185. doi:10.1111/pbi.12371

Wei, J., Man, Q., Ding, C., Hu, Y., Liu, M., Li, H., et al. (2019). Proteomic Investigations of Transcription Factors Critical in Geniposide-Mediated Suppression of Alcoholic Steatosis and in Overdose-Induced Hepatotoxicity on Liver in Rats. *J. Proteome Res.* 18, 3821–3830. doi:10.1021/acs.jproteome.9b00140

Wen, J., Jiang, F., Weng, Y., Sun, M., Shi, X., Zhou, Y., et al. (2019). Identification of Heat-Tolerance QTLs and High-Temperature Stress-Responsive Genes through Conventional QTL Mapping, QTL-Seq and RNA-Seq in Tomato. *BMC Plant Biol.* 19, 398. doi:10.1186/s12870-019-2008-3

Xu, Y. Q., Wei, G. Y., Zhou, Y., Ge, F., and Luo, G. M. (2014). Isolation and Characterization of Twenty-Two Polymorphic Microsatellite Markers from *Gardenia Jasminoides* (*Rubiaceae*). *J. Genet.* 94, e22–e24. doi:10.1007/s12041-014-0348-1

Xu, Z., Pu, X., Gao, R., Demurtas, O. C., Fleck, S. J., Richter, M., et al. (2020). Tandem Gene Duplications Drive Divergent Evolution of Caffeine and Crocin Biosynthetic Pathways in Plants. *BMC Biol.* 18, 63. doi:10.1186/s12915-020-00795-3

Yamakawa, H., Haque, E., Tanaka, M., Takagi, H., Asano, K., Shimosaka, E., et al. (2021). Polyploid QTL-seq towards Rapid Development of Tightly Linked DNA Markers for Potato and Sweetpotato Breeding through Whole-genome Resequencing. *Plant Biotechnol. J.* 19, 2040–2051. doi:10.1111/pbi.13633

Yan, H. H., Mudge, J., Kim, D.-J., Shoemaker, R. C., Cook, D. R., and Young, N. D. (2004). Comparative Physical Mapping Reveals Features of Microsynteny betweenGlycine max,Medicago Truncatula, andArabidopsis Thaliana. *Genome* 47, 141–155. doi:10.1139/g03-106

Yang, W., Wang, Y., Jiang, D., Tian, C., Zhu, C., Li, G., et al. (2020). ddRADseq-Assisted Construction of a High-Density SNP Genetic Map and QTL fine Mapping for Growth-Related Traits in the Spotted Scat (*Scatophagus argus*). *BMC genomics* 21, 278. doi:10.1186/s12864-020-6658-1

Yu, C., Yan, C., Liu, Y., Liu, Y., Jia, Y., Lavelle, D., et al. (2020). Upregulation of a *KN1* Homolog by Transposon Insertion Promotes Leafy Head Development in Lettuce. *Proc. Natl. Acad. Sci. USA* 117, 33668–33678. doi:10.1073/pnas.2019698117

Zhang, L., Guo, D., Guo, L., Guo, Q., Wang, H., and Hou, X. (2019). Construction of a High-Density Genetic Map and QTLs Mapping with GBS from the Interspecific F_1 Population of P. Ostii 'Fengdan Bai' and P. Suffruticosa 'Xin Riyuejin'. *Scientia Horticulturae* 246, 190–200. doi:10.1016/j.scienta.2018.10.039

Zhang, Z., Wei, T., Zhong, Y., Li, X., and Huang, J. (2016). Construction of a High-Density Genetic Map of *Ziziphus Jujuba* Mill. Using Genotyping by Sequencing Technology. *Tree Genet. Genomes* 12, 76. doi:10.1007/s11295-016-1032-9

Zhao, J., Li, H., Xu, Y., Yin, Y., Huang, T., Zhang, B., et al. (2021). A Consensus and Saturated Genetic Map Provides Insight into Genome Anchoring, Synteny of Solanaceae and Leaf- and Fruit-Related QTLs in Wolfberry (*Lycium* Linn). *BMC Plant Biol.* 21, 350. doi:10.1186/s12870-021-03115-1

Zhu, F., Saluja, M., Dharni, J. S., Paul, P., Sattler, S. E., Staswick, P., et al. (2021). PhenoImage : An Open-source Graphical User Interface for Plant Image Analysis. *Plant phenome j.* 4, e20015. doi:10.1002/ppj2.20015

Conflict of Interest: YX was employed by Adsen Biotechnology 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.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 05 January 2022
doi: 10.3389/fpls.2021.799482

Genetic Diversity and Population Structure of Sorghum [*Sorghum Bicolor* (L.) Moench] Accessions as Revealed by Single Nucleotide Polymorphism Markers

Muluken Enyew[1,2], Tileye Feyissa[1], Anders S. Carlsson[2], Kassahun Tesfaye[1,3], Cecilia Hammenhag[2] and Mulatu Geleta[2]*

[1] Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia, [2] Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma, Sweden, [3] Ethiopian Biotechnology Institute, Addis Ababa, Ethiopia

OPEN ACCESS

Edited by:
Andrés J. Cortés, Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

Reviewed by:
Zhenbin Hu, Saint Louis University, United States
Reza Darvishzadeh, Urmia University, Iran
Nemera Geleta Shargie, Agricultural Research Council of South Africa (ARC-SA), South Africa

***Correspondence:**
Muluken Enyew
muluken.birara@aau.edu.et;
mulukenbi@gmail.com;
muluken.birara.enyew@slu.se

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

Received: *21 October 2021*
Accepted: *03 December 2021*
Published: *05 January 2022*

Citation:
Enyew M, Feyissa T, Carlsson AS, Tesfaye K, Hammenhag C and Geleta M (2022) Genetic Diversity and Population Structure of Sorghum [Sorghum Bicolor (L.) Moench] Accessions as Revealed by Single Nucleotide Polymorphism Markers. Front. Plant Sci. 12:799482. doi: 10.3389/fpls.2021.799482

Ethiopia is the center of origin for sorghum [*Sorghum bicolor* (L.) Moench], where the distinct agro-ecological zones significantly contributed to the genetic diversity of the crops. A large number of sorghum landrace accessions have been conserved *ex situ*. Molecular characterization of this diverse germplasm can contribute to its efficient conservation and utilization in the breeding programs. This study aimed to investigate the genetic diversity of Ethiopian sorghum using gene-based single nucleotide polymorphism (SNP) markers. In total, 359 individuals representing 24 landrace accessions were genotyped using 3,001 SNP markers. The SNP markers had moderately high polymorphism information content (PIC = 0.24) and gene diversity (H = 0.29), on average. This study revealed 48 SNP loci that were significantly deviated from Hardy–Weinberg equilibrium with excess heterozygosity and 13 loci presumed to be under selection ($P < 0.01$). The analysis of molecular variance (AMOVA) determined that 35.5% of the total variation occurred within and 64.5% among the accessions. Similarly, significant differentiations were observed among geographic regions and peduncle shape-based groups. In the latter case, accessions with bent peduncles had higher genetic variation than those with erect peduncles. More alleles that are private were found in the eastern region than in the other regions of the country, suggesting a good *in situ* conservation status in the east. Cluster, principal coordinates (PCoA), and STRUCTURE analyses revealed distinct accession clusters. Hence, crossbreeding genotypes from different clusters and evaluating their progenies for desirable traits is advantageous. The exceptionally high heterozygosity observed in accession *SB4* and *SB21* from the western geographic region is an intriguing finding of this study, which merits further investigation.

Keywords: agro-ecological zone, genetic differentiation, geographical region, population structure, sorghum [*Sorghum bicolor* (L.) Moench]

INTRODUCTION

Sorghum [*Sorghum bicolor* (L.) Moench] is the fifth most important cereal crop in the world next to maize, rice, wheat, and barley in terms of both production and harvested area (FAOSTAT, 2019). It is a major food crop for more than 500 million people across Africa, Asia, and Latin America, particularly for those in the semi-arid tropical regions (Ejeta, 2005). It is grown in drought-prone areas where several other crops cannot reliably grow. Recent FAOSTAT data on annual global production of sorghum showed that it covered about 40 million ha of land and produced grains of ca 57.9 million metric tons (MMT) (FAOSTAT, 2019). The United States, Nigeria, and Ethiopia are the leading sorghum-producing countries in the world with a total production of 8.6, 6.7, and 5.2 MMT, respectively (Statista, 2020). In Africa, sorghum is the second most widely cultivated cereal crop, only surpassed by maize (FAOSTAT, 2019).

Ethiopia is considered as one of the centers of origin and diversity of sorghum (De Wet and Harlan, 1971) due to the presence of wild relatives and diversified forms of the crop in the country. The sorghum gene pool in the country has been used as novel sources of germplasm for crop improvements. For example, genotypes harboring genes that confer resistance to ergot and green bug (Wu et al., 2006) as well as high lysine (Singh and Axtell, 1973) and drought-tolerant (Borrell et al., 2000) sorghum genotypes were identified from the Ethiopian accessions.

Studying the genetic diversity of a crop is very important for effective germplasm management, utilization, and genotype selection for crop improvement (Bucheyeki et al., 2009). It is the most important step for conserving and increasing the rate of genetic gain in crop-breeding programs. The level of genetic diversity within a species is commonly used to measure the level of species adaptability and survival in unpredictable environmental conditions (Rao and Hodgkin, 2002; Govindaraj et al., 2015). Similarly, the level of genetic variation within a population is the basis for germplasm selection in plant breeding and is vital for crop improvement (Mohammadi and Prasanna, 2003). Hence, the conservation and utilization of plant genetic variation are crucial to human food security (Rao and Hodgkin, 2002).

Sorghum is a predominantly self-pollinated diploid species (Poehlman and Sleper, 1979) with $2n = 2\times = 20$ chromosomes. It has a small genome relative to other cereal crops, which is about 730 Mbp (Paterson et al., 2009). Its whole genome was sequenced and made accessible for public use[1] (Paterson et al., 2009; McCormick et al., 2018), which facilitated the development of DNA markers, such as single nucleotide polymorphism (SNPs) for various applications, including analyses of population genetics and identification of genomic regions associated with complex traits through quantitative trait loci (QTL) and association mapping (Too et al., 2018; Girma et al., 2019).

The genetic diversity of crop species can be studied through morphological, biochemical, and molecular markers (Rao et al., 1996; Geleta and Labuschagne, 2005; Mehmood et al., 2008; Enyew et al., 2021). Previous studies on the genetic diversity of sorghum have been carried out by using random amplified polymorphism DNA (RAPD) analysis (Ayana et al., 2000; Ruiz-Chután et al., 2019), simple sequence repeat (SSR) markers (Djè et al., 2000; Ghebru et al., 2002; Manzelli et al., 2007; Ali et al., 2008; Wang et al., 2009; Ng'uni et al., 2011, 2012; Burow et al., 2012; Adugna et al., 2013; Adugna, 2014; Mofokeng et al., 2014; Motlhaodi et al., 2014, 2017), and express sequence tags (EST) SSR markers (Ramu et al., 2013), SNP markers (Cuevas et al., 2017; Afolayan et al., 2019; Cuevas and Prom, 2020). More recently, a few studies have been performed on the genetic diversity of Ethiopian sorghum accessions using SNP markers (Girma et al., 2019; Menamo et al., 2021; Wondimu et al., 2021). These studies brought out the contribution of geographic regions and agro-ecological zones for the genetic variation and population structure of sorghum grown in Ethiopia. However, these studies did not consider genetic variation within populations, as the analyses were based on either a single plant per accession or a pool of individual plants treated as a single sample per accession). Ethiopian Biodiversity Institute (EBI) has conserved more than 9,432 sorghum accessions collected from diverse agro-ecologies across the country.[2] However, the genetic diversity of most of the accessions in the collection remains molecularly uncharacterized. Therefore, this study analyzes the genetic diversity and population structure of selected Ethiopian sorghum accessions using SNP markers in order to generate highly important information, which together with previous research results, lead to deeper insight on the sorghum gene pool in the country and beyond.

MATERIALS AND METHODS

Plant Materials

Twenty-four Ethiopian sorghum landrace accessions originally collected by the EBI were obtained from Melkassa Agricultural Research Center (MARC). The accessions were selected to represent three agro-ecological zones according to the classification by Amede et al. (2015) viz. cool/subhumid, cool/semiarid, and warm/semiarid zones (**Supplementary Figure 1**). **Supplementary Table 1** provides details about these accessions, including the sampling locations, as well as major morphological and phenological characteristics. Photographs showing panicle diversity in the Ethiopian sorghum that represents these accessions are provided as **Supplementary Figure 2**.

Planting, Sampling, and Genomic DNA Extraction

Sorghum seeds representing the 24 accessions were planted using plastic pots filled with soil in a greenhouse at the Department of Plant Breeding, SLU, Sweden. Two weeks after planting, the leaf tissues from individual plants were collected using a sample collection kit provided by LGC-Genomics (Berlin, Germany), as described by Tsehay et al. (2020). Each accession was represented by 15 individual plants, except accession *SB10*, which was represented by 14 individuals; hence 359 genotypes were sampled

[1] http://genome.jgi-psf.org/Sorbi1/Sorbi1.info.html

[2] https://www.ebi.gov.et/biodiversity/conservation/genetic-material-holdings/

in total. The samples were then sent to LGC Genomics (Berlin, Germany) where genomic DNA extraction was conducted for subsequent genotyping. High-quality genomic DNA, suitable for next-generation sequencing (NGS), was extracted using the Sbeadex plant kit.[3]

SNP Selection, Assay Design, Sequencing, and Genotype Calling

The vast majority of SNPs (97%) used in this study were selected from sorghum genome SNP database SorGSD,[4] a web-portal that provides genome-wide SNP markers for diverse sorghum genetic resources (Luo et al., 2016). Among different sorghum lines in the database, *Cherekit* (an Ethiopian sorghum landrace accession) was targeted for selecting the SNPs. For genotyping, SeqSNP method (an advanced NGS method for genotyping target SNPs) was used. Initially, 12,316 SNPs were targeted for high-specificity (without allowing for off-target hit) assay design, using *Sorghum bicolor* v3.1.1 genome in Phytozome 12.1[5] as a reference (Paterson et al., 2009; McCormick et al., 2018). Additionally, 380 SNPs within functionally annotated sorghum genes were identified through the Basic Local Alignment Search Tool (BLAST), searching the genes targeting *S. bicolor* v3.1.1 genome sequence using Phytozome 12.1 search function were targeted for the assay design. Out of the total 12,696 targets used for the high specificity assay design, 9,495 were totally covered (two oligo probes per target), 1,631 were partially covered (one oligo probe), whereas 1,190 failed.

For the seqSNP genotyping, 5,000 SNPs were selected among the totally covered SNPs, based on their distribution across the sorghum genome. The number of SNPs targeted on chromosome-1 to chromosome-10 included in the order, 532, 521, 572, 506, 497, 515, 437, 446, 465, and 509 (refer to **Supplementary Table 2**). One hundred fifty-seven of these SNPs belong to 51 functionally annotated genes (**Supplementary Table 3**). This was followed by the construction of SeqSNP kit LGC, Biosearch Technologies (Berlin, Germany) comprising 10,000 high-specificity oligo probes for the 5,000 target SNPs and construction of a sequencing library. The target sequencing was conducted using Illumina NextSeq 500/550 v2 system with 75 bp single read sequencing mode. In the end, ca 973,000 reads per sample were obtained and the effective target of SNP coverage per sample was 175 times on average. After sequencing, the reads were adapter-clipped and quality-trimmed to get a minimum Phred quality score of 30 over a window of ten bases. After discarding reads shorter than 65 bases. the quality trimmed reads were aligned against the reference genome using Bowtie2 v2.2.3 (Langmead and Salzberg, 2012), and the SNP genotyping pipeline was set to diploid genotyping with a minimum coverage of eight reads per sample per locus. The variant identification and genotype calling were done using Freebayes v1.0.2-16 (Garrison and Marth, 2012).

[3]https://biosearch-cdn.azureedge.net/assetsv6/sbeadex-plant-data-sheet.pdf

[4]http://sorgsd.big.ac.cn

[5]www.phytozome.net

Data Analysis

The site frequency spectra were analyzed for each accession using DnaSP version 6 (Rozas et al., 2003). The nucleotide diversity (Nei, 1987) and Tajima's D (Tajima, 1989) were calculated using the PopGenome package (Pfeifer et al., 2014) in R software (R Core Team) to reveal the genome-wide pattern of variation using a sliding window approach (window size = 1 Mb, step size = 200 kb), in line with previous studies in sorghum (Yan et al., 2018), maize, and common bean (Lai et al., 2010; Cortés and Blair, 2018).

The mean effective number of alleles (Ne), observed heterozygosity (Ho), expected heterozygosity (He), Shannon information index (I), and gene flow (Nm) for each SNP marker and accession were estimated using GenAlEx 6.5 (Peakall and Smouse, 2012), and the gene diversity (H) and the polymorphism information content (PIC) were performed using PowerMarker (Liu and Muse, 2005). The Hardy–Weinberg equilibrium (HWE) test was also done using GenAlEx 6.5.

Analysis of Molecular Variance (AMOVA) within and among the accessions as well as at higher hierarchical levels were done using the software, Arlequin ver. 3.5.2.2 (Excoffier and Lischer, 2010). Arlequin was also used for estimating pairwise genetic differentiation between accessions and groups and for detecting outlier SNP markers through a non-hierarchical finite island model. The significance of the differentiation of accessions and groups was tested by 10,000 permutations. The joint distribution of population differentiation (F_{ST}) and heterozygosity (heterozygosity within populations)/(1 – F_{ST}) were obtained according to Excoffier and Lischer (2010). The loci under selection were identified based on the F_{ST} significance level of $P < 0.01$.

The principal coordinates analysis (PCoA) was done using GenAlEx 6.5. The bootstrap-supported unweighted pair group method with arithmetic mean (UPGMA) clustering based on Nei's genetic distance (Nei and Takezaki, 1983) was performed using PowerMarker v 3.25 (Liu and Muse, 2005) and the resulting trees were visualized using MEGA-X (Kumar et al., 2018). STRUCTURE v. 2.3.4 software (Pritchard et al., 2000) was used for the Bayesian clustering of the 359 individuals representing the 24 sorghum accessions, at the burn-in period length of 100,000 and a Markov Chain Monte Carlo (MCMC) replications of 100,000. The structure analysis was done for K ranging from two to ten, with ten iterations at each K, to determine the optimum number of clusters (genetic populations). The optimum K value was predicted following the simulation method of Evanno et al. (2005) using STRUCTURE HARVESTER version 0.6.92 (Earl, 2012). A bar plot for the optimum K was determined through Clumpak beta version (Kopelman et al., 2015).

RESULTS

The Quality and Level of Polymorphism of SNP Markers

In the data matrix of 5,000 SNP loci for the 359 sorghum genotypes, missing data accounted for 2.8% and, of those with data, 94.1% were homozygous. Among the 5,000 target SNP loci,

4,301 (86%) were polymorphic, whereas 699 loci (14%) were monomorphic across the 359 genotypes. Among the 4,301 SNP loci, 4,256, 42, and 1 were bi-, tri- and tetra-allelic, respectively, whereas two loci had the combination of SNP and length variants. Tri- and tetra-allelic loci were excluded from further analysis. The filtering of the 4,256 bi-allelic SNP data, based on the percentage of missing data and minimum allele frequency (MAF) resulted in different numbers of loci varying from 2,259 loci with less than1% missing and greater than10% MAF to 3,089 loci with less than 5% missing and greater than 5% MAF. For the population genetics analyses, 3,001 bi-allelic SNP loci with missing data less than 2% and MAF greater than 5% were used (**Supplementary Figure 3**). About 26% of the markers had MAF between 5 and 10% whereas 31% had MAF between 11 and 20%. About 40% of the markers had MAF greater than 2% (**Supplementary Figure 3** and **Supplementary Table 4**). The SNP markers had a moderately balanced distribution across the chromosomes, ranging from 252 SNPs (8.4%) on chromosome 8–371 SNPs (12.4%) on chromosome 3 (**Supplementary Table 4**).

The site frequency spectrum revealed high variation in the MAF distribution of the SNPs among the 24 sorghum landrace accessions (**Figure 1**). All individuals in 67% of the accessions (horizontally, the first 16 accessions in **Figure 1**) had major alleles in most of their SNP loci although to a different extent. Interestingly, one individual from each of the remaining eight accessions carried minor alleles across most of their SNP loci. The expected site-frequency spectrum determined using a coalescent approach matched the observed frequency distributions fairly well for the accessions, *SB2*, *SB7*, and *SB19* while it was inversely related to the observed frequency distributionsin accession *SB4* (**Figure 1**). The genome-wide diversity across sorghum landrace accessions was quantified using a sliding window approach (window size = 1 Mbp, step size = 200 kb) to explore the genomic signatures of diversity in sorghum. The analyses resulted in an overall average nucleotide diversity (π) and Tajima's D of 1.2 and 1.4/Mb, respectively (**Figure 2**). It is clear from **Figure 2** that SNPs representing the centromere regions of each chromosome almost do not exist among the 3,001 SNPs used in this study, and thus the diversity estimates were extremely low or zero. At chromosome level, the highest average nucleotide diversity and Tajima's D were recorded in chromosome 9 (1.5 and 2.2/Mb) and the lowest in chromosome 7 (1.0/Mb) and chromosome 8 (0.7/Mb), respectively (**Figure 2**). Previously reported candidate loci for domestication are found at the genomic regions with notably low diversity on chromosomes 2, 4, and 7 (**Figure 2**). The effective number of alleles found across the 3,001 SNP markers ranged from 1.01 to 1.98 with a mean of 1.16. Observed heterozygosity (Ho) varied from 0.0 to 0.96 with a mean of 0.06 while the mean expected heterozygosity (He) was 0.10 with individual values per locus ranging from 0.01 to 0.49. Similarly, the gene diversity estimates per locus varied from 0.10 to 0.50 with a mean value of 0.29 (**Figure 3** and **Supplementary Table 4**). The average PIC of the loci was 0.24 with individual values ranging from 0.09 to 0.37. In the case of fixation indices, the minimum, maximum, and mean values for F_{IS} were –0.96, 1.00, and 0.45, for F_{IT} they were –0.93, 1.00, and 0.79, and for F_{ST}, they were 0.01, 0.95, and 0.63, respectively. The estimates of gene flow (Nm) per locus showed wide variation, ranging from 0.01 to 20.23, with a mean of 0.20 (**Figure 3** and **Supplementary Table 4**).

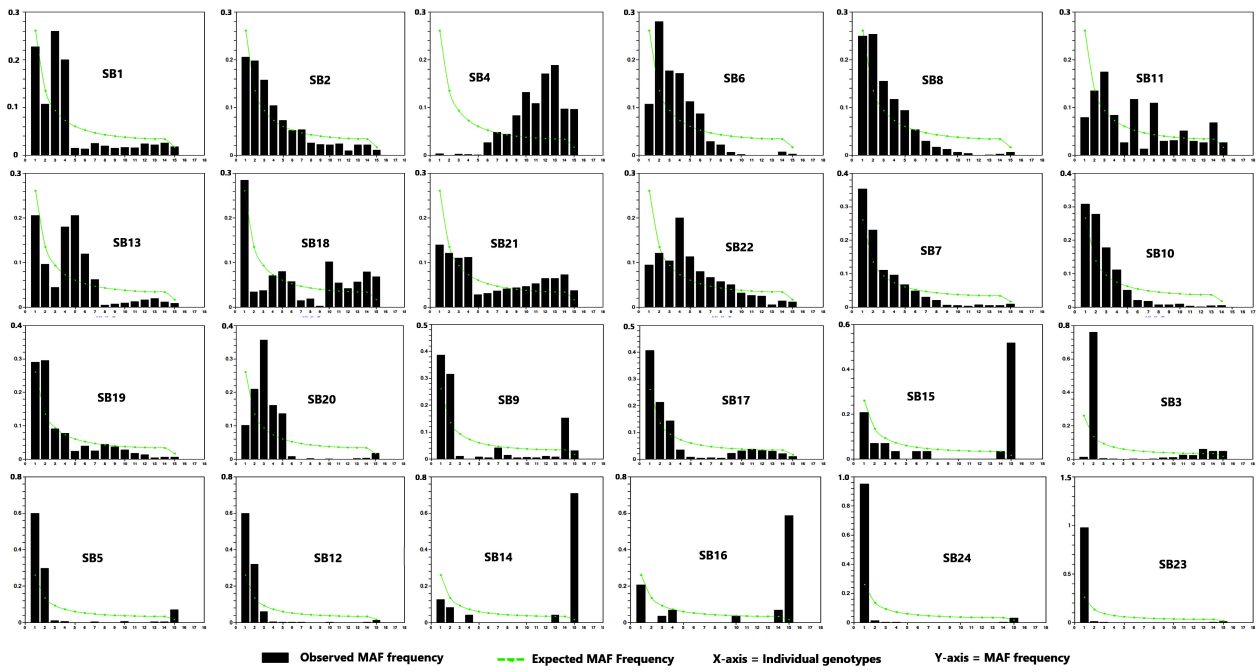


FIGURE 1 | The pattern of site frequency spectrum based on the proportion of the minor allele frequency (MAF) of single nucleotide polymorphism (SNP) in the 24 sorghum accessions.

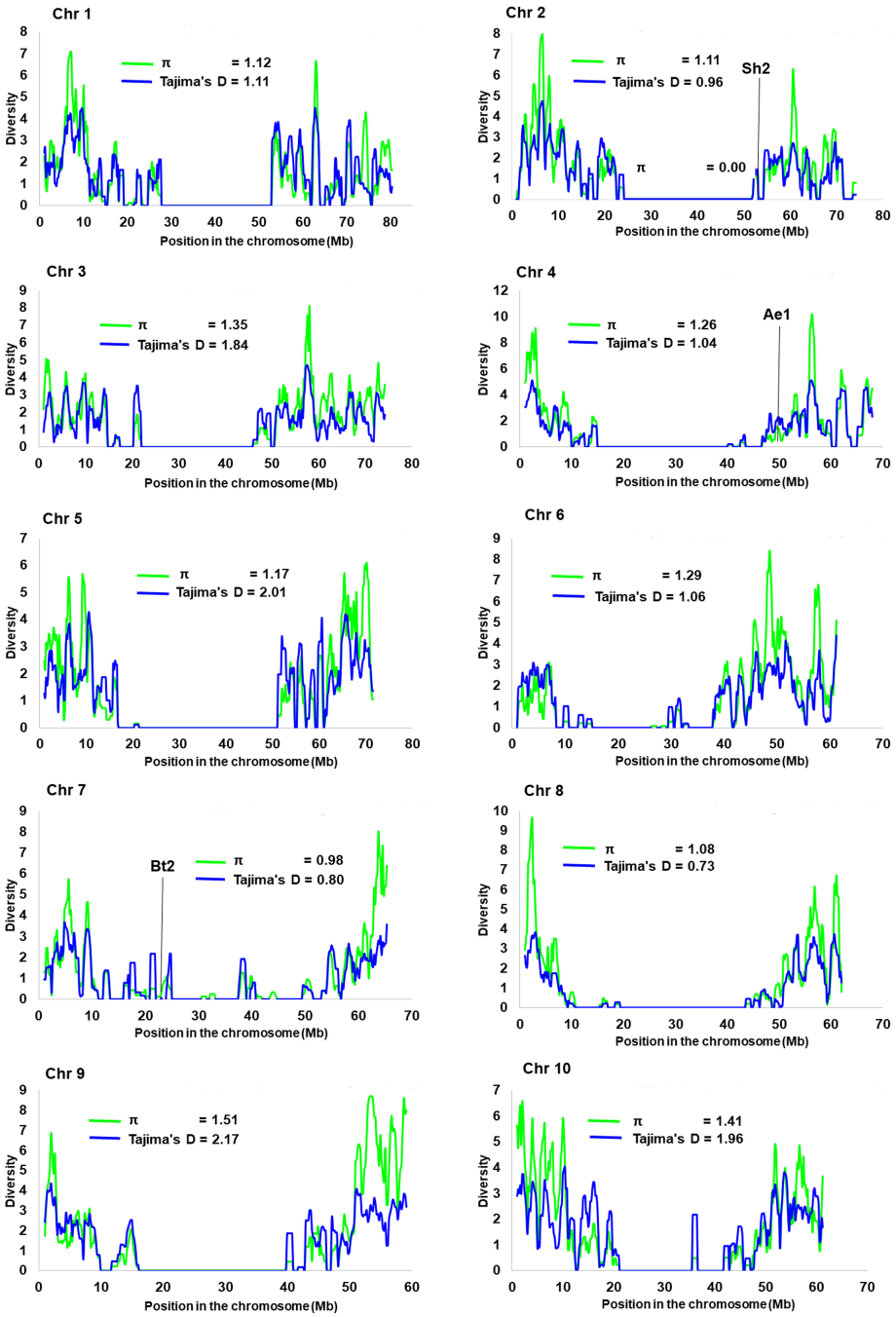
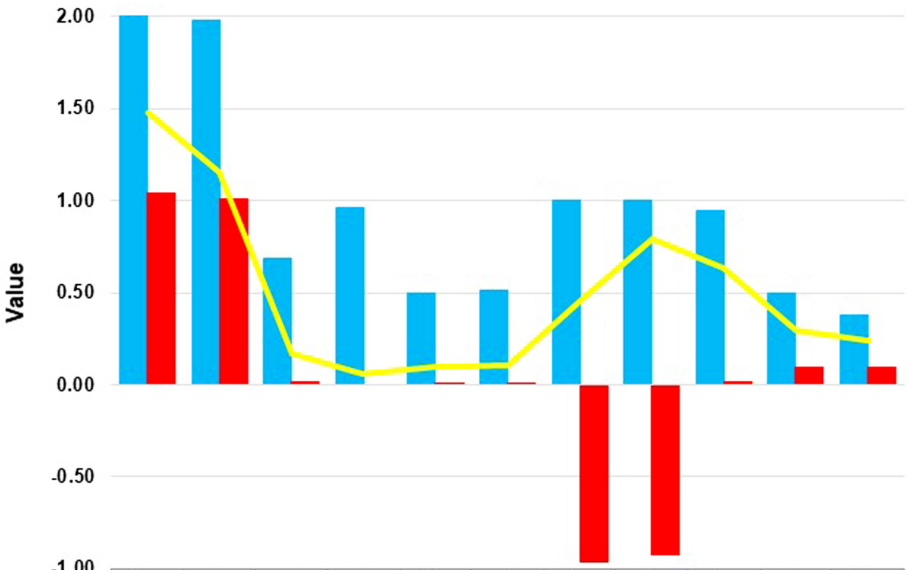


FIGURE 2 | Genome-wide pattern of diversity in the 359 individual plants representing the 24 sorghum accessions. A sliding window approach (window size = 1 Mb, step size = 200 kb) was used to analyze nucleotide diversity and Tajima's D. The black vertical lines on chromosomes 2, 4, and 7 show the positions of *shrunken2* (sh2), *amylose extender1* (ae1), and *brittle2* (bt2), respectively, which were previously identified as domestication loci in maize and sorghum that are localized at regions of low diversity. The overall average nucleotide diversity (π) and Tajima's D were 1.2/Mb and 1.4/Mb, respectively.

Based on the HWE test, 99.5% of the SNP markers showed significant deviation from HWE (**Supplementary Table 4**). Among the 3,001 SNP loci, 97.9% were heterozygote-deficient, whereas 1.6% (48 loci) had excess heterozygosity showing significant deviation from HWE ($P < 0.05$). The candidate genes containing SNP markers showing excess heterozygosity and their annotated functions were retrieved from SorGSD[6] and further evaluated. Among the 48 SNP loci that showed excess heterozygosity, nine SNPs lacked one of the three possible genotypes expected in a bi-allelic polymorphic locus under the

[6]https://ngdc.cncb.ac.cn/sorgsd/



	Na	Ne	I	Ho	He	uHe	Fis	Fit	Fst	H	PIC
Max	2.00	1.98	0.69	0.96	0.49	0.51	1.00	1.00	0.95	0.50	0.38
Min	1.04	1.01	0.02	0.00	0.01	0.01	-0.96	-0.93	0.01	0.10	0.09
Mean	1.48	1.16	0.17	0.06	0.10	0.11	0.45	0.79	0.63	0.29	0.24

FIGURE 3 | The mean, minimum (Min), and maximum (Max) values for the number of allele (Na), number of effective allele (Ne), Shannon informative index (I), observed heterozygosity (Ho), unbiased expected heterozygosity (uHe), expected heterozygosity (He), fixation indices (Fis, Fit, Fst), polymorphic information content (PIC), and gene diversity (H) for the 3001 polymorphic SNP loci.

assumption of HWE. The change in amino acid sequences of the corresponding genes was obtained in all SNPs, except three SNPs (snp_sb001000020838, snp_sb042060612417, and snp_sb042061102446) (**Supplementary Table 5**).

Genetic Diversity Analysis

The 3,001 polymorphic SNP markers revealed a wide range of variation in the Ethiopian sorghum germplasm, as estimated using different population genetics parameters across the 24 accessions, and are summarized in **Table 1**. The effective number of alleles of the accessions varied from 1.01 to 1.46 with a mean of 1.21, whereas the mean Shannon's Information index (I) was 0.25 with individual values ranging from 0.0 (*SB14* and *SB15*) to 0.42 (*SB21*). The lowest and the highest Ho values varied from 0.01 (*SB5*, *SB14,* and *SB15* and *SB16*) to 0.25 (*SB21*) with a mean of 0.07. Likewise, the He and unbiased expected heterozygosity (uHe) of the accessions ranged from 0.0 to 0.27 and 0.0 to 0.28, respectively with a mean of 0.15 (**Table 1**). The lowest values were recorded in accessions, *SB14, SB15,* and *SB16,* whereas *SB21* recorded the highest values for these parameters. The percent polymorphic loci (PPL) of the accessions varied from 0.8 to 91.4% with a mean of 47.7%. The fixation index (F) showed wide variation with values ranging from –0.76 (*SB14*) to 0.84 (*SB3*). Overall, accession *SB21* showed the highest value for Ne, Ho, He, uHe, I, and the number of locally common alleles (NLCA) and PPL while *SB14*, *SB15*, and *SB16* showed the lowest values for all genetic diversity parameters analyzed (**Table 1**).

Among the agro-ecological zones, warm/semiarid zones showed the highest values for Ne, Ho, He, uHe, and I whereas cool/subhumid zones showed the lowest in the majority of the genetic diversity estimates. Among the groups of accessions in the three agro-ecological zones, the highest value of PPL, which is equal to 99% and the number of private allele per locus were recorded in warm/semiarid and cool/semiarid zones (**Table 1** and **Supplementary Table 6**). In terms of geographic regions, accessions from the western geographic regions showed the highest values in most of the genetic diversity parameters analyzed (I, Ho, He, and uHe) (**Figure 4** and **Table 1**). For example, the eastern, northern, and western accessions had uHe values of 0.24, 0.21, and 0.37, respectively. Accessions from the eastern geographic region showed the highest value PPL, which is equal to 99.7% and in the number of private alleles. Four private alleles were recorded for the eastern region with MAF ranging from 0.19 to 0.47 whereas two and one private alleles were detected in the accessions originated from the western and northern regions, respectively (**Table 1** and **Supplementary Table 6**). With regard to peduncle shape, accessions with bent peduncles were more diverse than those with erect peduncles as shown by the values of I, Ho He, uHe, and PPL (**Figure 4** and **Table 1**). One hundred thirty-nine alleles were specific to accessions with bent peduncles shape with MAF ranging from 0.09 to 0.32, whereas only one private allele with MAF of 0.14 was specific to accessions with erect peduncles (**Supplementary Table 6**).

TABLE 1 | Summary of different genetic diversity estimates based on 3,001 SNP markers for each of the 24 sorghum accessions and for a group of accessions grouped according to different agro-ecological zones (cool/semiarid, cool/subhumid, and warm/semiarid), geographical regions (eastern, northern and western), and peduncle shape (bent and erect).

Accession	Na	Ne	I	Ho	He	uHe	F	PPL	NPA	NLCA
SB1	1.806	1.248	0.282	0.064	0.169	0.174	0.494	80.6	0.0	0.008
SB2	1.663	1.214	0.239	0.080	0.144	0.149	0.344	66.3	0.0	0.006
SB3	1.317	1.101	0.108	0.023	0.065	0.067	0.837	31.7	0.0	0.003
SB4	1.459	1.394	0.298	0.157	0.210	0.217	0.255	45.9	0.0	0.006
SB5	1.104	1.018	0.023	0.013	0.013	0.013	0.196	10.4	0.0	0.002
SB6	1.832	1.221	0.284	0.059	0.165	0.171	0.607	83.2	0.0	0.008
SB7	1.484	1.110	0.142	0.043	0.081	0.083	0.380	48.4	0.0	0.004
SB8	1.740	1.175	0.229	0.090	0.131	0.135	0.262	74.0	0.0	0.006
SB9	1.292	1.095	0.095	0.019	0.058	0.060	0.503	29.2	0.0	0.003
SB10	1.741	1.161	0.217	0.072	0.122	0.126	0.322	74.1	0.0	0.007
SB11	1.427	1.200	0.192	0.039	0.123	0.128	0.557	42.7	0.0	0.003
SB12	1.433	1.051	0.086	0.031	0.043	0.044	0.242	43.3	0.0	0.005
SB13	1.692	1.235	0.265	0.056	0.162	0.167	0.531	69.2	0.0	0.005
SB14	1.008	1.006	0.005	0.006	0.003	0.003	−0.757	0.8	0.0	0.001
SB15	1.010	1.006	0.005	0.006	0.003	0.003	−0.589	1.0	0.0	0.002
SB16	1.010	1.007	0.005	0.007	0.004	0.004	−0.685	1.0	0.0	0.001
SB17	1.791	1.214	0.239	0.074	0.141	0.146	0.283	79.1	0.0	0.007
SB18	1.504	1.253	0.224	0.065	0.147	0.152	0.368	50.4	0.0	0.003
SB19	1.568	1.153	0.180	0.031	0.106	0.109	0.558	56.8	0.0	0.005
SB20	1.335	1.082	0.110	0.046	0.063	0.065	0.263	33.5	0.0	0.003
SB21	1.914	1.457	0.416	0.253	0.271	0.280	0.038	91.4	0.0	0.008
SB22	1.680	1.271	0.287	0.071	0.179	0.185	0.536	68.0	0.0	0.006
SB23	1.410	1.034	0.063	0.032	0.029	0.030	−0.041	41.0	0.0	0.004
SB24	1.232	1.023	0.038	0.023	0.019	0.019	−0.057	23.2	0.0	0.002
Mean	1.754	1.209	0.250	0.067	0.147	0.152	0.419	47.7	0.0	0.005
cool/semiarid	1.994	1.434	0.419	0.043	0.268	0.269	0.782	99.4	5.0	0.000
cool/subhumid	1.909	1.387	0.381	0.051	0.243	0.244	0.700	90.9	0.0	0.000
warm/semiarid	1.995	1.519	0.482	0.089	0.316	0.318	0.660	99.5	5.0	0.000
Mean	1.966	1.447	0.427	0.061	0.276	0.277	0.715	96.6	3.3	0.000
eastern	1.997	1.378	0.389	0.053	0.243	0.244	0.715	99.7	4.0	0.000
northern	1.844	1.291	0.276	0.023	0.176	0.177	0.678	89.8	1.0	0.000
western	1.779	1.272	0.273	0.045	0.170	0.171	0.628	97.4	2.0	0.000
Mean	1.974	1.647	0.535	0.158	0.364	0.369	0.514	95.6	2.3	0.000
bent	2.000	1.564	0.518	0.080	0.341	0.342	0.760	99.9	139.0	0.000
erect	1.954	1.336	0.323	0.030	0.206	0.206	0.650	95.3	1.0	0.000
Mean	1.977	1.450	0.420	0.050	0.274	0.274	0.710	97.6	70.0	0.000

Na = Number of different alleles; Ne = effective number of alleles; I = Shannon's information index; Ho = observed heterozygosity; He = expected heterozygosity; uHe = unbiased expected heterozygosity; F = fixation index; PPL = percent polymorphic loci; NPA = number of private alleles; NLCA = number of locally common alleles found in 25% or fewer accessions or group of accessions.

Genetic Differentiation of Accessions and Hierarchical Groups

The results of the AMOVA without grouping the accessions showed that 64.5% of the total variation was observed among accessions and 35.5% within accessions (F_{ST} = 0.65; F_{IS} = 0.47, $P < 0.001$) (**Table 2**). Additionally, hierarchical AMOVA was conducted by grouping the accessions according to their geographic regions, agro-ecological zones of their collection sites, and their peduncle shape. In this analysis, 19.5% of the total variation was observed among the geographical regions, which is a highly significant differentiation (F_{CT} = 0.20, $P < 0.001$). Similarly, significant differentiation was found among peduncle shape groups with 4.3% of the total variation between them (F_{CT} = 0.04, $P < 0.05$) (**Table 2**). However, only 1.83% of the total variation accounted for the variation among the agro-ecological zones, which is statistically insignificant (F_{CT} = 0.02 and P = 0.17) (**Table 2**).

Population Differentiation and Gene Flow

The pairwise population differentiation analysis revealed significant differentiation among all pairs of accessions with F_{ST} values ranging from 0.18 to 0.99 (**Figure 5** and **Supplementary Table 7**) except in the case of *SB16* vs. *SB12*, which was not significant (F_{ST} = 0.02, $P > 0.05$). The pairs of accessions with

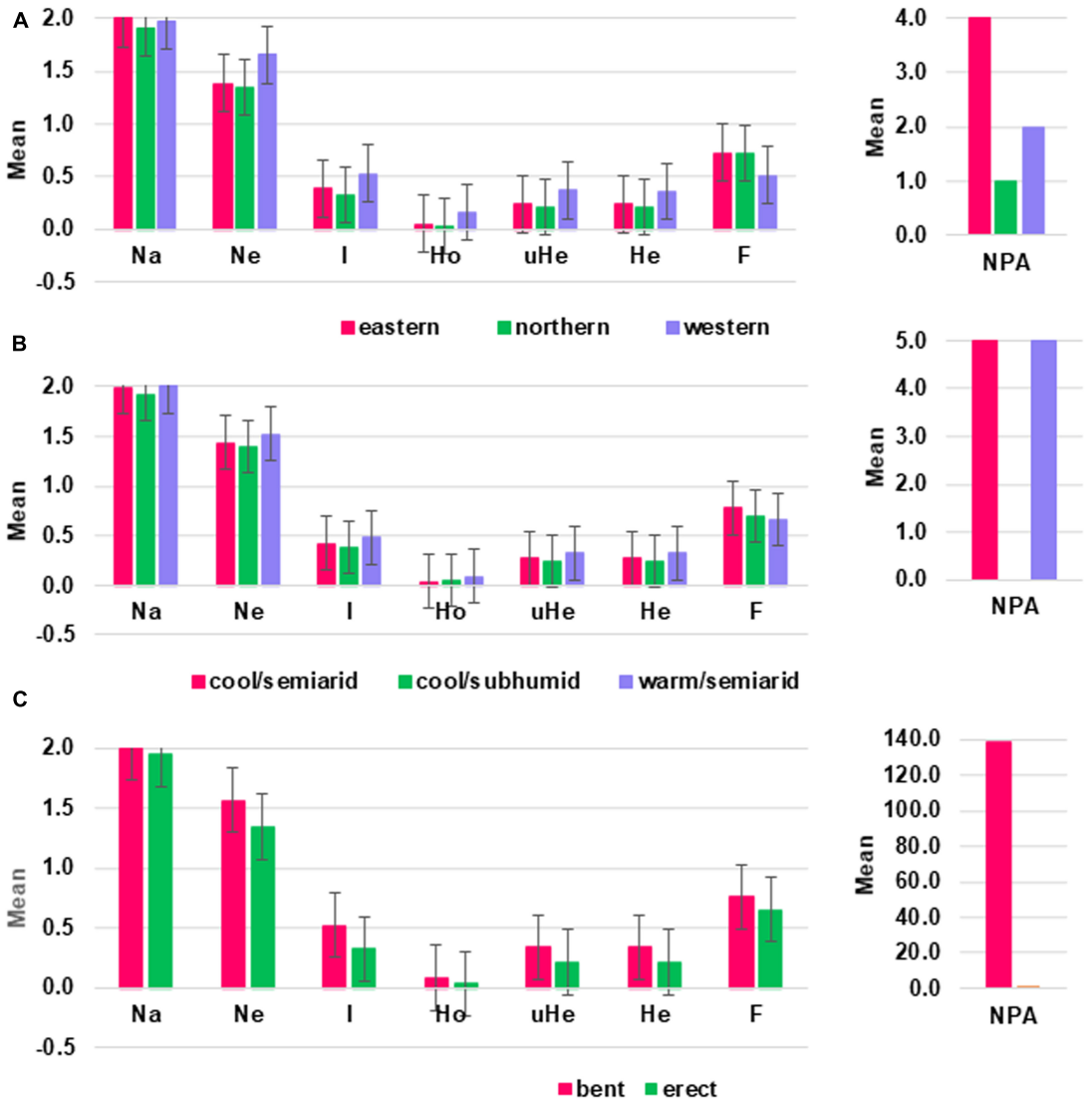


FIGURE 4 | Graphs displaying mean values of different genetic diversity parameters estimated based on 3,001 SNP markers for a group of sorghum accessions grouped according to their **(A)** geographic regions, **(B)** agro-ecological zones, and **(C)** peduncle shape. Na = No. of different alleles; Ne = effective number of alleles; I = Shannon's information index; Ho = observed heterozygosity; uHe = unbiased expected heterozygosity; He = expected heterozygosity; F = fixation index; NPA = number of private alleles.

the highest F_{ST} value (0.99) were *SB14* vs. *SB15* and *SB15* vs. *SB16,* corresponding to the lowest estimate of gene flow (Nm = 0; **Supplementary Table 7**). The mean F_{ST} values for the differentiation of each accession from all other accessions varied from 0.47 to 0.81. Accessions *SB15*, *SB14*, and *SB5* were the most differentiated with F_{ST} values of 0.81, 0.80, and 0.78, respectively, whereas *SB18* was the least differentiated accession (F_{ST} = 0.47) (**Figure 5** and **Supplementary Table 7**).

The analyses of the average number of pairwise differences (π_{xy}) and the pairwise net number of allele differences (Nei's distance, *d*) between the accessions revealed a wide variation with π_{xy} ranging from 1.2 (*SB12* vs. *SB16*) to 1,197.2 (*SB6* vs. *SB15*) and *d* ranging from 0.001 (*SB12* v.s *SB16*) to 0.56 (*SB21* vs. *SB6* and SB1 vs. *SB14*) (**Figure 6** and **Supplementary Table 8**). Accessions *SB1*, *SB6,* and *SB21* also showed a higher pairwise net number of allele differences (Nei's distance, *d*) with other accessions (**Figure 6** and **Supplementary Table 8**). In line with the results of the pairwise F_{ST} analysis, the average number of pairwise differences and Nei's distance were the lowest for *SB16* vs. *SB12* suggesting that these two accessions are genetically very similar. The average number of pairwise differences within accessions also showed wide variation with the values ranging from 10 (*SB14*) to 840 (*SB21*). This parameter was very low for *SB15* and *SB16*, as with *SB14* (**Figure 6** and **Supplementary Table 8**).

At the geographic region level, the pairwise F_{ST} values among each pair of the three groups were significant ($P < 0.001$). However, accessions from the western region were the most

TABLE 2 | Analysis of molecular variance (AMOVA) for 24 accessions without grouping, and by grouping them based on their geographic regions, agro-ecological zones, and peduncle shapes.

Source of variation	DF	SS	Variance components	Percentage of variation	Fixation indices	Probability (*P*) value
Among accessions	23	206414.2	292.1 Va	64.5	F_{ST} = 0.65	Va and F_{ST} < 0.001
Among individuals within accessions	335	79341.1	76.0 Vb	16.8	F_{IS} = 0.47	Vb and F_{IS} < 0.001
Within individuals	359	30426.5	84.8Vc	18.7	F_{IT} = 0.81	Vc and F_{IT} < 0.001
Total	717	316181.8	452.9			
Among geographic regions	2	55111.0	95.5 Va	19.52	F_{CT} = 0.20	Va and F_{CT} < 0.001
Among accessions within geographic regions	21	151303.1	235.6 Vb	48.15	F_{SC} = 0.60	Vb and F_{SC} < 0.001
within accessions	694	109767.6	158.2 Vc	32.33	F_{ST} = 0.68	Vc and F_{ST} < 0.001
Total	717	316181.8	489.2			
Among agro-ecological zones	2	21393.9	8.3 Va	1.83	F_{CT} = 0.02	Va and F_{CT} = 0.17
Among accessions within agro-ecological zones	21	185020.3	289.3 Vb	63.47	F_{SC} = 0.65	Vb and F_{SC} < 0.001
within accessions	694	109767.6	158.7 Vc	34.70	F_{ST} = 0.65	Vc and F_{ST} < 0.001
Total	717	316181.8	455.8			
Among peduncle shape groups	1	15762.6	19.9 Va	4.30	F_{CT} = 0.04	Va and F_{CT} < 0.05
Among accessions within peduncle shape groups	22	190651.5	284.4 Vb	61.50	F_{SC} = 0.64	Vb and F_{SC} < 0.001
Within accessions	694	109767.6	158.17 Vc	34.20	F_{ST} = 0.66	Vc and F_{ST} < 0.001
Total	717	316181.8	462.4			

*The 24 accessions were grouped into four geographic regions, three agro-ecological groups based on world classifications (Amede et al., 2015), and two peduncle shape groups (**Supplementary Table 1**). PS, peduncle shape; DF, degrees of freedom; SS, sum of square; Va, Vb, and Vc, variance explained by the source of variation; F_{ST}, F_{IS}, F_{IT}, F_{CT} and F_{SC}, fixation indices.*

distinct with higher differentiation from those from the northern and eastern regions (F_{ST} = 0.40 and 0.35, respectively). Among the three pairs, accessions from the eastern vs. northern regions were the least differentiated (F_{ST} = 0.12) (**Supplementary Table 7**). Similar to that of geographic regions, the F_{ST} values among each pair of the accessions from the three agro-ecological zones were also significant ($P < 0.001$). The accessions belonging to the cool/subhumid group were the most differentiated having F_{ST} values of 0.13 and 0.10 against warm/semiarid and cool/semiarid groups, respectively. The warm/semiarid vs. cool/semiarid groups were the least differentiated (F_{ST} = 0.08) among the three pairs (**Supplementary Table 7**). The average number of pairwise differences and Nei's distance among the geographic regions and agro-ecological zones had a similar pattern with that of pairwise F_{ST}-based differentiation, revealing that the western region was the most differentiated group. In the case of pairwise differences within regions, accessions from the western region had the highest variation whereas the lowest was recorded for the northern region. With regard to agro-ecological zones, warm/semiarid and cool/subhumid zones showed the highest and lowest variations, respectively (**Supplementary Table 7**).

The non-hierarchical finite island model-based analysis involving the examination of the joint distribution of F_{ST} and heterozygosity among accessions to detect loci under selection revealed 74 SNP loci that were highly significant ($P < 0.01$). Among them, 61 loci had low F_{ST} value (ranging from –0.02 to 0.35), and hence were considered as candidates for balanced selection. Whereas 13 loci (**Table 3**) had high F_{ST} values (ranging from 0.81 to 0.94), and hence considered as under directional selection. The MAF of these loci ranged from 0.05 to 0.34. The markers were distributed on chromosomes 1, 5, 6, 7, and 9 with over 50% of them located on chromosome 7 (**Table 3**). The candidate genes containing these SNP markers and their putative functions were identified through BLAST searching the sorghum v3.1 genome at Phytozome 12.1 (**Table 3**).

Cluster Analyses of Individual Genotypes and Accessions

The unweighted pair group method with arithmetic mean-based cluster analysis of the 359 individual genotypes generated a dendrogram of three major clusters, which were denoted by different line colors in **Figure 7**. The cluster analysis at the accession level resulted in the clustering of the 24 accessions into two groups. In the case of individual genotypes, Cluster I consisted of 316 individuals, whereas Cluster II comprised 43 individuals, respectively. The cluster analysis showed that at least the majority of individuals from the same accessions were clustered together (**Figure 7**). All individuals of an accession were clustered closely together in several cases. For example, all individuals from accessions, *SB21* and *SB4* were clustered in Cluster I and Cluster II, respectively. In other cases, a few individuals of an accession were placed under different clusters. For instance, two individuals from accession *SB1*, one individual from *SB6*, *SB17*, and *SB22* were separated from the other members of their accession and grouped with other genotypes in different clusters. Except in a few cases, most accessions were clearly clustered based on their geographic regions (**Figure 8A**). On the other hand, the clustering pattern of the accessions according to their agro-ecological zones or administrative regions

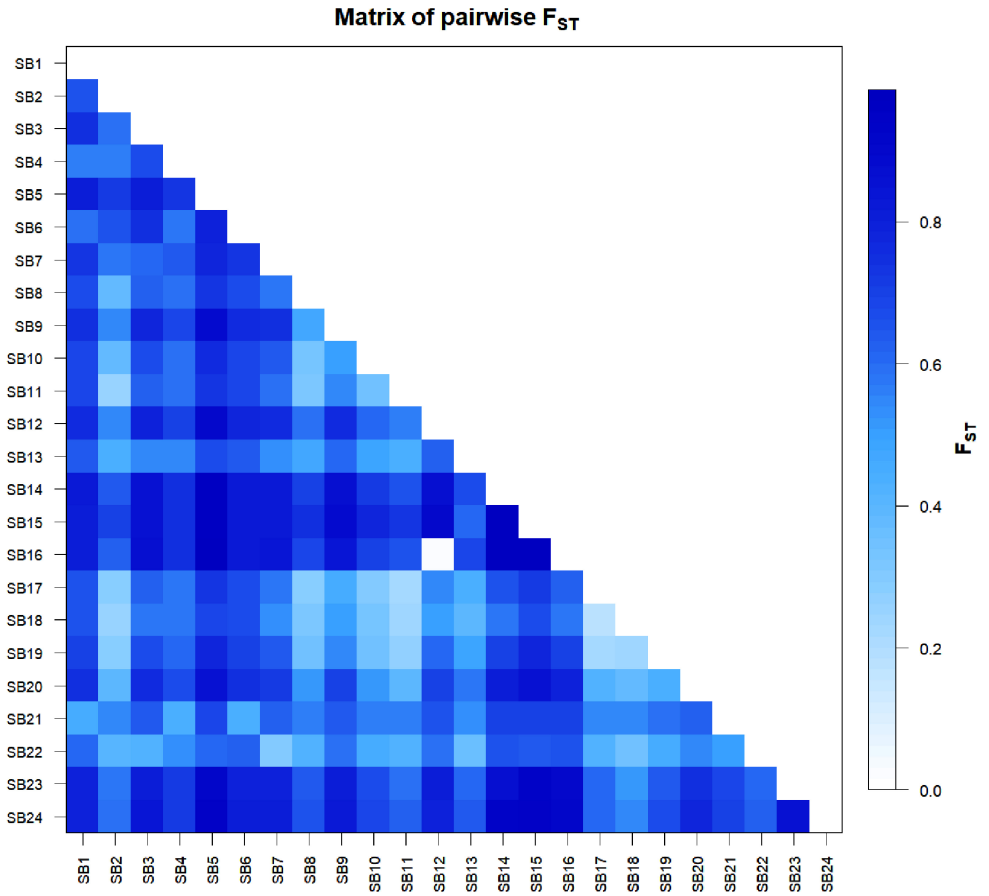


FIGURE 5 | Graphical display of pairwise genetic differentiation (F_{ST}) among the 24 sorghum accessions. The differentiation between each pair was significant ($P < 0.05$) except in the case of SB12 vs. SB16.

was less resolved, as accessions were mostly clustered irrespective of their groups (**Figures 8B,C**).

Principal Coordinate Analysis

Principal coordinate analysis was performed to determine the relationship between the sorghum accessions and individuals within the accession, which grouped the accessions into three separate clusters (**Figure 9A** and **Supplementary Figure 4**). The first and second coordinates explained 29.5 and 12.4% of the total variation among the accessions, respectively. Similar to the cluster analysis, PCoA revealed that accessions *SB1*, *SB4*, *SB6*, and *SB21* are the most differentiated groups being clearly separated from the other accessions along the first principal coordinate (**Figure 9A**).

Population Structure

The admixture model-based population structure of the 359 individuals representing the 24 accessions was inferred using STRUCTURE software. The analysis of the STRUCTURE output using STRUCTURE HARVESTER program (Earl, 2012) that implemented ΔK method of Evanno et al. (2005) revealed that the optimal number of genetic clusters is two (**Supplementary Figure 5**). The results suggest that the 24 sorghum accessions originate from two genetic populations as graphically depicted in **Figure 9B**. In line with the results of the cluster analysis and PCoA, accessions *SB1, SB4, SB6,* and SB21 were significantly differentiated groups, as the majority of their alleles belong to a different genetic population (represented by orange in **Figure 9B**) as compared to the other accessions.

DISCUSSION

SNP Markers and Their Use in Genetic Diversity Analysis of Sorghum Gene Pool

Genetic diversity analysis of crop species is an important step in detecting alleles that could be used for their improvement through breeding. The Ethiopian sorghum gene pool has been used as a novel source of biotic and abiotic stress tolerance, greatly contributing to the improvement of sorghum, globally (Adugna, 2014). The gene pool has been utilized in various studies that aimed at the identification of novel QTLs and genes governing complex traits (Cuevas and Prom, 2013, 2020; Cuevas et al., 2017; Menamo et al., 2021). Since polymorphism within

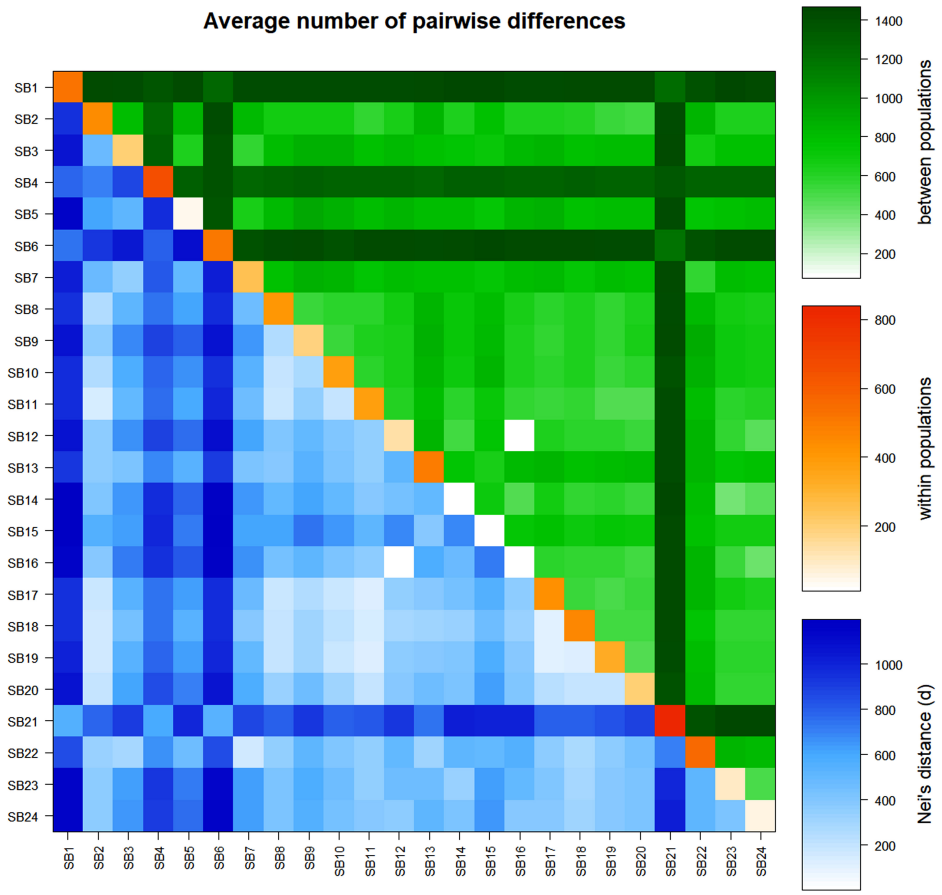


FIGURE 6 | Average number of pairwise differences within and between sorghum accessions: average number of pairwise differences among the accessions (π_{xy}) (above diagonal, in green), average number of pairwise differences within accessions (π) (diagonal, in orange), and pairwise net number of allele differences among the accessions (d) (below diagonal, in blue).

genes or their close vicinity is expected to be the main basis of phenotypic variation, priority was given to SNPs located in genes in the SNP selection process in this study. Because of simplicity and abundance in plant genomes, bi-allelic SNPs are the most commonly used SNPs used in genetic analyses. In the present study, 86% of the genotyped bi-allelic SNP loci were polymorphic, which can be considered high. This is most likely because, the SNP selection was mainly made based on the SNPs recorded for the Ethiopian sorghum genotype, *Cherekit* at the SorGSD database. The vast majority of the SNP loci (94.1%) were homozygous across the 359 individual samples genotyped, which is not surprising as sorghum is a self-pollinating crop.

The variation in allele frequency distribution among accessions shown by the analysis of site frequency spectrum indicates a high level of genetic diversity in the Ethiopian sorghum. Accessions containing individual genotypes dominated by minor alleles across the loci require further investigations to reveal the phenotypic diversity of desirable traits. The overall nucleotide diversity (π) of 1.2 recorded in this study is in agreement with the result of a previous study on sorghum landraces (Mace et al., 2013). However, it is higher than the values reported in some other studies on sorghum (Morris et al., 2013; Yan et al., 2018). Similarly, the overall Tajima's D value recorded in this study was 1.4, which is lower and higher than values reported in Morris et al. (2013) and (Mace et al., 2013), respectively. Among the seven starch-related genes, *amylose extender1* (*ae1*), *brittle2* (*bt2*), *Opaque2* (*O2*), *shrunken1* (*sh1*), *shrunken2* (*sh2*), *sugary1* (*su1*), and w*axy1* (*wx1*), previously identified as candidates of domestication loci (Whitt et al., 2002; De Alencar Figueiredo et al., 2010; Morris et al., 2013), three of them (*sh2*, *ae1,* and *bt2*) are found at the genomic regions with notably low diversity on chromosomes 2, 4, and 7, respectively (**Figure 2**). The *bt2* gene on chromosome 7 coding for a starch biosynthesis enzyme has been shown to be a likely domestication locus in sorghum and maize (Whitt et al., 2002; De Alencar Figueiredo et al., 2010; Morris et al., 2013). The low recombination rates in the pericentromeric region or the presence of other loci under selection in this region may be the reason for the low diversity in the present study and previous studies on sorghum (Morris et al., 2013).

The average Ho of 0.06 obtained in the present study was in line with the results of previous studies on sorghum employing

TABLE 3 | The list of 13 SNP loci that were identified as loci under selection and their descriptions.

SNP markers	Chr.	SNP Pos.	Ref/Alt	MAF	Het	F_{ST}	Map pos	Candidate Gene	Annotation
snp_sb001000600993	6	4719375	T/A	0.33	0.30	0.91	4719016.4719660	Sobic.006G026400	No annotated domain for this protein
snp_sb001000665443	6	30979222	G/A	0.34	0.27	0.89	30971365.30982488	Sobic.006G044400	Similar to OSIGBa0097I24.1 protein
snp_sb042060260697	1	67217203	T/C	0.17	0.46	0.83	67215996.67219771	Sobic.001G384700	Similar to Zinc finger C-x8-C-x5-C-x3-H type familyprotein, expressed
snp_sb042060594735	5	61927575	C/T	0.29	0.35	0.94	61927471.61932993	Sobic.005G150200	Weakly similar to Putative uncharacterized protein
snp_sb042060764024	7	5101345	A/G	0.18	0.45	0.84	5098301.5104103	Sobic.007G050400	Similar to DEAD-box ATP-dependent RNA helicase 42
snp_sb042060834108	7	54216799	C/A	0.34	0.28	0.88	54216242.54222782	Sobic.007G127600	Similar to Os02g0653400 protein
snp_sb042060834531	7	54578827	G/C	0.23	0.41	0.84	54578165.54580032	Sobic.007G129300	Weakly s.t Putative uncharacterized protein
snp_sb042060855091	7	65318268	C/G	0.13	0.49	0.91	65311844.65318939	Sobic.007G225800	Similar to Proliferating-cell nucleolar antigen-like protein
snp_sb042060855137	7	65355302	A/C	0.13	0.49	0.91	65354308.65356452	Sobic.007G226300	Similar to Putative uncharacterized protein
snp_sb042060855138	7	65357757	T/C	0.13	0.49	0.91	65356668.65359547	Sobic.007G226400	Similar to Pentatricopeptide (PPR) repeat-containing protein-like
snp_sb042060855361	7	65431483	C/T	0.13	0.49	0.91	65424095.65433354	Sobic.007G227100	Similar to Os08g0482100 protein
snp_sb042061021321	9	52527630	C/T	0.14	0.48	0.83	52527443.52529178	Sobic.009G169300	Weakly similar to Os05g0470900 protein
snp_sb042061023058	9	53254749	T/A	0.05	0.51	0.81	53254243.53259288	Sobic.009G177600	K10683 - BRCA1-associated RING domain protein 1

Chr., chromosome; SNP Pos., SNP position in the corresponding sorghum chromosome; Ref/Alt, reference and alternative alleles; MAF, minor allelic frequency; Het, heterozygosity; F_{ST}, population differentiation; Map Pos., map positions of the candidate genes in the corresponding sorghum chromosome.

SNP markers (Cuevas et al., 2017) and SSR markers (Ng'uni et al., 2011) (Ho = 0.04), (Ramu et al., 2013; Motlhaodi et al., 2014) (Ho = 0.09), (Motlhaodi et al., 2017) (Ho = 0.03). The Ho was expected, as sorghum is a predominantly a self-pollinating crop (Poehlman and Sleper, 1979). Gene diversity (H) and PIC are the most common measures of polymorphism of markers, which shed light on the evolutionary pressure on the alleles and the mutation rate at a locus over time (Shete et al., 2000; Wilkinson et al., 2012). The total genetic diversity in a population can be estimated through the analyses of a large number of informative markers across their genome (Melchiorre et al., 2013). The gene diversity of the SNP markers across all accessions in this study ranged from 0.1 to 0.50 with a mean of 0.29, which is high. Informative markers could be used for genotyping populations for genetic diversity studies, and the informativeness of the markers can be measured by their PIC value (Salem and Sallam, 2016). In the case of bi-allelic SNP markers, the maximum PIC value of 0.375 is attained when both alleles have a frequency of 0.5. In the present bi-allelic SNP-based study, the PIC values ranged from 0.09 to 0.375 with the overall average of 0.24, which is comparable with previous studies on sorghum using SNP markers (Afolayan et al., 2019; Silva et al., 2021; Wondimu et al., 2021). Forty-seven percent of these SNP loci have a PIC value of greater than 0.25 and hence they are highly informative and could be used for various applications including population genetic studies of sorghum.

Selections, both natural and artificial, as well as inbreeding, contribute to the deviation of populations from HWE. In this study, 98% of the loci showed heterozygote deficiency while 1.60% of the loci showed excess heterozygosity. Since sorghum is a predominately self-pollinating species, heterozygote deficiency at the vast majority of the loci can be attributed to inbreeding. However, the small proportion of loci showing excess heterozygosity suggests that they could be under selection or linked to loci under selection. Among the SNP loci that showed excess heterozygosity, nine loci lacked one of the two homozygous genotypes. The data suggest that one of the two alleles at each locus reduces the fitness of homozygous genotypes, or the locus is linked to another locus within its corresponding gene or the nearby gene that has a significant fitness value. Most of these SNP markers are within the coding region of genes. For instance, snp_sb001000687053, snp_sb001000723312, snp_sb042060543510, snp_sb042060515233, and snp_sb042060517985 are within the coding region of senescence-related gene 1, tetratricopeptide repeat (TPR)-like superfamily protein, cysteine proteinases superfamily protein, C-terminal domain phosphatase-like 4, and hydroxyproline-rich glycoprotein family protein, respectively. These genes had major roles in the growth, development, physiology, and biotic and abiotic stress tolerances in plants. For instance, hydroxyproline-rich glycoproteins (HRGPs) play a major role in the growth and development of plants (Showalter et al., 2016) while cysteine proteinases play an important physiological process ranging from seed germination (Becker et al., 1994) to senescence (Valpuesta et al., 1995). Therefore, further study that investigates the effect of these SNPs on

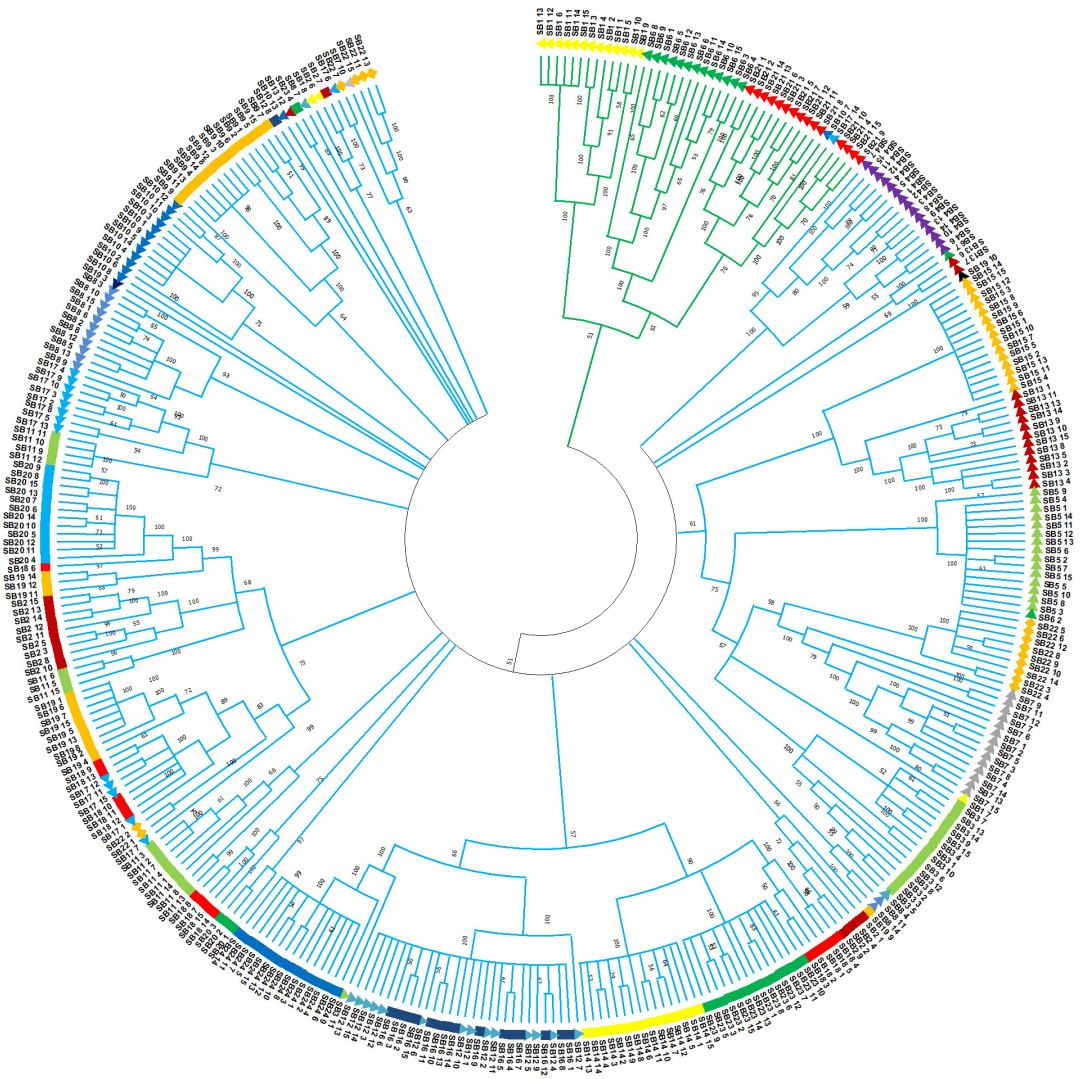


FIGURE 7 | Unweighted pair group method with arithmetic mean (UPGMA) dendrogram of 359 individuals representing the 24 sorghum accessions generated based on Nei's genetic distance (Nei and Takezaki, 1983). The individual samples were coded in a way that the first two letters (SB) with either two- or three-digit numbers represent their accessions and the last two-digit numbers represent the codes for the individual plant in that accession. Individuals denoted by the same color and shape belong to the same accession.

the response of sorghum to abiotic and biotic stresses is of high significance.

Genetic Diversity Within Accessions

The average He (0.15), I (0.25), and PPL (47.7%) obtained in the present study suggest low genetic variation within the sorghum accessions. In general, the relatively low genetic variation within landrace accessions in the present and previous studies on sorghum (Ng'uni et al., 2011; Motlhaodi et al., 2017) is likely due to the combination of its inbreeding nature and due to the strict selection criteria of farmers. However, the variation within accessions varied widely. In this regard, accessions from the western region [Benishangul-Gumuz, Gambella, and Southern Nations, Nationalities and Peoples' Region (SNNPR)] had higher variation than other accessions with *SB21* being the most diverse accession followed by *SB4*. Accessions from this region (*SB21, SB1,* and *SB4*) are characterized by bent peduncle and light brown seeds with the exception of the red seed color of *SB1*.

Mengistu et al. (2020) also reported higher gene diversity and PIC for accessions from the Benshangul-Gumuz, Gambella, and SNNP regions as compared to the other regions in Ethiopia. The higher variation within accessions from these regions may suggest less human selection pressure on the landraces as compared to sorghum grown elsewhere in the country. Since the genetic diversity of populations implies their potential to adapt to environmental changes (Markert et al., 2010), sorghum landraces from this region may serve as a potential source of genes for biotic and abiotic stresses. Another interesting result of this study is a significantly higher Ho in two of the three accessions (*SB4* and *SB21*) from the western regions as compared to all other

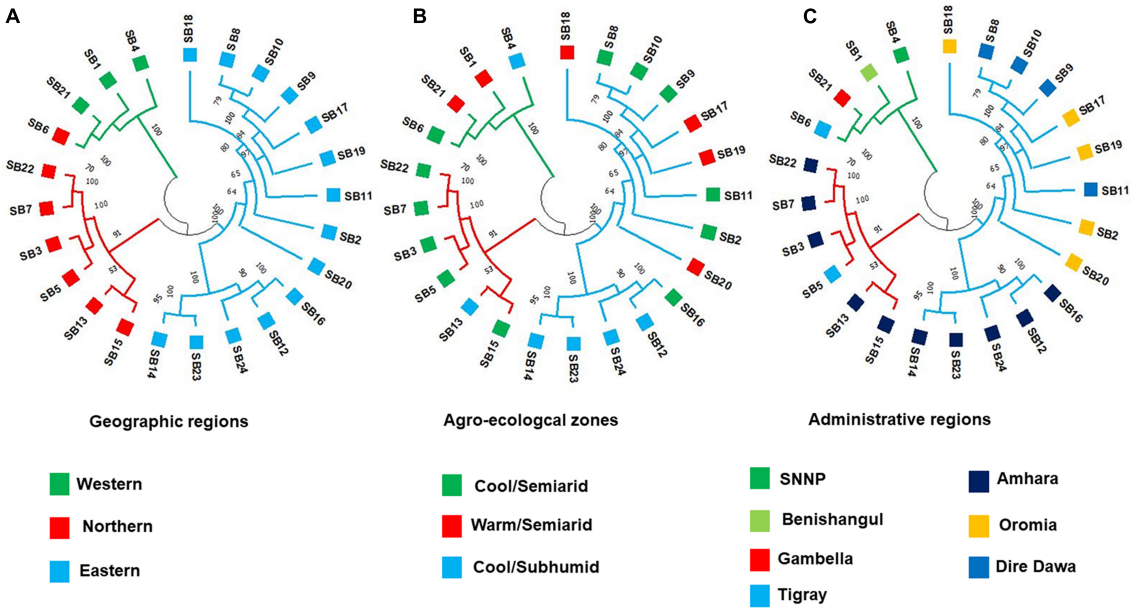


FIGURE 8 | The UPGMA dendrogram of the 24 accessions generated based on Nei's genetic distance (Nei and Takezaki, 1983) calculated using the genotypic data of 3,001 SNP markers. Accessions denoted by the same color labels belong to the same **(A)** geographic region, **(B)** agro-ecological zone, and **(C)** administrative region.

accessions. Higher Ho suggests a higher outcrossing rate in these accessions, which might have allowed for gene flow through pollen and hence increased the variation within the accessions. The results suggest the western region as an important source of sorghum genotypes with desirable traits, such as tolerance to biotic and abiotic stresses. On the other hand, most accessions from Northern Ethiopia (Tigray and Amhara) had very low variation within accessions. Their average Ho was 0.03, indicating that the vast majority of the loci in the genotypes of these accessions were homozygous. In this group, accessions *SB5*, *SB14*, *SB15*, and *SB16* can be regarded as pure lines, as individuals within each accession are almost identical across the whole loci. On the other hand, other accessions in this group (*SB6*, *SB13*, and *SB22*) are more diverse although their heterozygosity is still very low. Since the loss of heterozygosity increases the chance of deleterious recessive alleles being expressed in the progeny (Radosavljević et al., 2015), these accessions may be more susceptible to biotic or abiotic stresses unless they have been selected for tolerance against these stresses over time.

Genetic Differentiation of Accessions and Hierarchical Groups

In this study, most of the total variations (64.5%) were observed among the accessions than within the accessions (35.5%). The lower genetic variation within the accessions is expected in self-pollinating crops like sorghum (Hamrick, 1983). In addition, strict farmers' selection for crop improvement might have contributed to the lower within-accession variation, which were clearly displayed in accessions, such as *SB14* and *SB15*. Previous genetic diversity studies through SNP and SSR markers also showed a higher genetic variation among sorghum accessions than within the accessions. For instance, SNP-based genetic diversity study on sorghum accessions from Ethiopia showed that the variation among and within the accessions accounted for 59.6 and 40.4% of the total variation, respectively (Mengistu et al., 2020). Similarly, genetic diversity study through SSR markers on sorghum accessions from Zambia revealed 82 and 18% genetic variations among and within the accession variations, respectively (Ng'uni et al., 2011). Motlhaodi et al. (2017) reported a significant genetic variation among 22 accessions of sorghum, which accounted for 66.9% of the total variation while the within accession variation accounted for 23.6%. However, high genetic variation within sorghum accessions were reported on sorghum studied through SNP markers (Afolayan et al., 2019) and SSR markers (Manzelli et al., 2007; Adugna, 2014), suggesting that the accessions are not under selection processes.

Several studies have shown that the diversity of sorghum is associated with geography, agro-ecology, ethnicity, or botanical racial classifications (Barnaud et al., 2007; Ng'uni et al., 2011; Faye et al., 2019; Menamo et al., 2021). Significant genetic variations among the geographic regions and peduncle shape groups were observed in this study as shown by hierarchical AMOVA. Among the geographic regions, the western and eastern regions had higher genetic diversity than the northern region as shown by average He and the percentage of polymorphic loci, which were higher than the overall average (He = 0.24 and PIC = 89%). The western region accessions were the most distinct, with higher differentiation from those of the northern and eastern geographic regions. The major sorghum growing area (northern region) of the country had relatively low genetic variation probably due to intensive farmers' selection of landraces to cope with the local

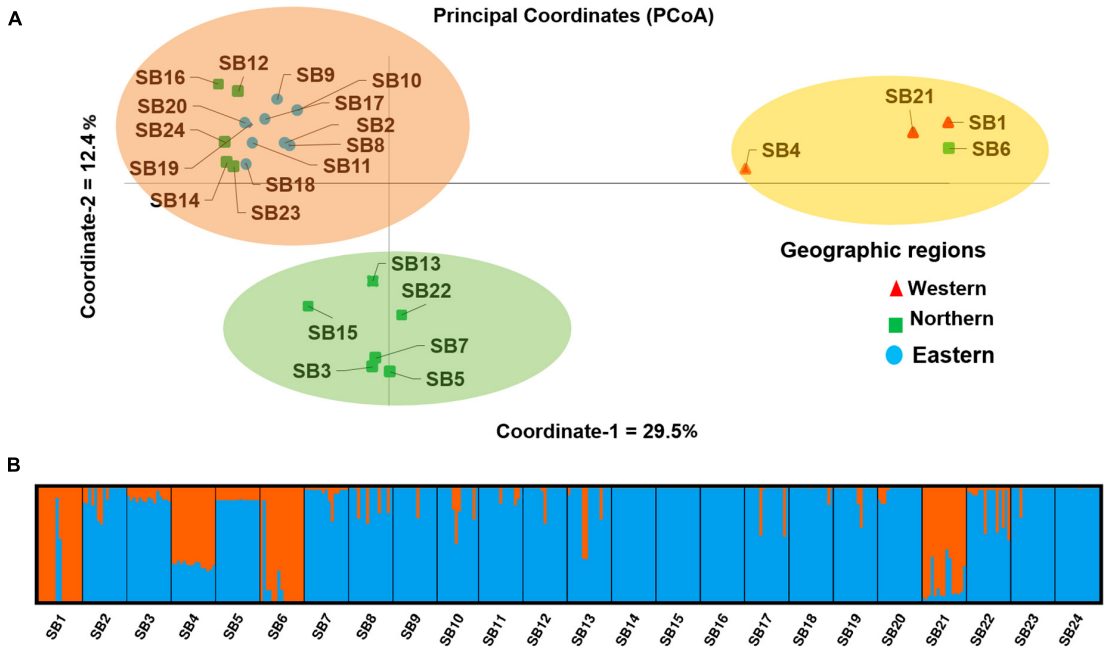


FIGURE 9 | (A) Principal coordinates analysis (PCoA) showing the clustering pattern of the 24 Ethiopian sorghum accessions and accessions denoted by the same color labels and shapes belonging to the same geographic region and **(B)** a graphical display of the population genetic structure of the 24 sorghum accessions for K = 2. The two colors represent the two clusters (genetic populations) and each color of an accession represents the average proportion of the alleles that placed that accession under the corresponding clusters.

environmental factors, such as the duration of the rainy season. The diversity of the crop has been reduced over time due to the recurrent drought in this major sorghum-growing region of the country. Overall, farmers in the drought-prone lowland areas tend to use early maturing and high yielding types and or shift their production systems to more vulnerable and low yielding early maturing crop species, such as tef (*Eragrostis tef*) (Adugna, 2014), which may provide genetic erosion of the sorghum landraces in these regions. High adoption of early maturing improved varieties in drought-prone areas in the northern region was also reported (Tesfaye et al., 2013).

Private alleles represent a unique genetic variability at certain loci of a particular population or hierarchically grouped populations. In this study, private alleles were not detected at the population level, but were recorded in all geographic regions. The Eastern region had a higher number of private alleles as compared to the western and northern regions, and hence it may serve as a rich source of desirable alleles for sorghum improvement. Private alleles generally support the potential to respond to a selection or have evolutionary significance (Petit et al., 1998). Information on private alleles is crucial for selecting highly diverse genotypes that can be used in breeding programs as a source of parental lines for crossbreeding that would eventually lead to new cultivars enriched with desirable alleles (Brondani et al., 2006; De Oliveira Borba et al., 2009; Salem and Sallam, 2016). The presence of more private alleles in the eastern region suggests the good *in situ* conservation status of sorghum in that location. Hence, further studies that explore the region for highly desirable traits need to be conducted, especially for use in sorghum-breeding programs.

Sorghum genotypes showed a significant genetic differentiation based on their peduncle shape, possibly because the shape of the peduncle influences the mating system, with the architecture of very bent peduncle obstructing pollination with outcrossed pollen. A more interesting finding was that accessions with bent peduncles exhibited higher genetic variation on average than those with erect peduncles. Unlike previous studies on Ethiopian sorghum (Menamo et al., 2021; Wondimu et al., 2021), sorghum accessions were not significantly differentiated according to agro-ecology in this study. However, a high significant genetic difference among the three pairs of agro-ecological zones was observed and the warm/semiarid zones showed the highest genetic diversity among the agro-ecological zones. Private alleles were detected from warm/semiarid and cool semiarid zones. Cool/subhumid zones, however, did not exhibit any private allele.

In the present study, 13 SNP loci were identified as loci under selection through the determination of the joint distribution of F_{ST} and heterozygosity. More than 50% of these SNP loci are located on chromosome 7 of the sorghum genome, suggesting that this chromosome carries many genes under natural selection or targeted by farmers directly or indirectly during and after domestication. These SNPs include those located in genes coding for zinc finger CCCH type family protein, DEAD-box ATP-dependent RNA helicase 42 and pentatricopeptide (PPR) repeat-containing proteins, which play a crucial role in plant responses

to biotic and abiotic stresses (Peng et al., 2012; Xing et al., 2018; Nidumukkala et al., 2019). Hence, further study on these loci using individual genotypes that carry different alleles may shed more light on their significance in terms of desirable traits.

The Clustering Pattern and Population Structure of the Sorghum Accessions

Unweighted pair group method with arithmetic mean clustering based on Nei's genetic distance (Nei and Takezaki, 1983) placed the individuals from the 24 accessions into three clusters. In line with the generally low genetic variation within accessions revealed through different analyses, there was a clear clustering pattern of individual genotypes according to their accessions. At the accession level, the cluster analysis generated three distinct clusters that matched the three clusters of the PCoA, which explained 42% of the total variation in its first two principal axes. The STRUCTURE analysis also generally agrees with the observed clustering pattern although it suggested two genetic populations (K = 2) as the best representation of the germplasm studied. Most of the alleles of the most distinct clusters in UPGMA and PCoA analyses (containing *SB1*, *SB4*, *SB6,* and *SB21*) originate from the first genetic cluster of STRUCTURE analysis (shown orange in **Figure 9**). Hence, it is interesting to crossbreed individual genotypes in these accessions with genotypes of genetically uniform accessions (e.g., *SB14* and *SB15*), and evaluate the progeny generations for desirable traits.

In this study, the significant differentiation among geographic groups but not among agro-ecological groups revealed through AMOVA was also evident in the cluster analysis at the level of accessions. Based on redundancy analysis in their recent study on sorghum, Menamo et al. (2021) reported that agro-ecology is more important than the administrative region in defining the genetic variation in sorghum, which is not in agreement with the present study. The present study showed that the genetic diversity of Ethiopian sorghum landrace accessions was more structured along the geographical regions than along the administrative regions or agro-ecological zones. The lack of clear genetic differentiation of sorghum along the administrative regions, which was also previously reported (Ayana and Bekele, 2000; Desmae et al., 2016; Wondimu et al., 2021), could be explained by a high gene flow because of extensive exchange of seeds among farmers across adjacent regions where sorghum is a major crop.

CONCLUSION

In this study, SeqSNP method was used to genotype diverse sorghum accessions using a combination of previously developed and newly identified gene-based SNP markers. Despite the fact that they were gene-based, the SNP markers revealed a comparable genetic variation from the previous studies using SNP markers in sorghum. About half of the SNP markers can be regarded as highly informative and can be prioritized for future population genetics studies. A significant number of loci exhibited excess heterozygosity and/or were presumed to be under selection, some of which are located within genes playing crucial roles in plant responses to biotic and abiotic stresses. Further research on these loci using genotypes carrying different alleles may shed light on their significance in terms of desirable traits. The observed highly significant genetic differentiation among the sorghum accessions will be beneficial to the sorghum breeders in selecting desirable parents for crossbreeding. The sorghum accessions formed three distinct clusters, and it is therefore interesting to crossbreed genotypes from different clusters to evaluate their progeny for desirable traits. In this study, highly significant variations were observed among the geographic regions and peduncle-shaped groups. Compared to the western and northern regions, the eastern region had a higher number of private alleles, and hence it may serve as a rich source of desirable alleles for improving sorghum. Lastly, given that sorghum is generally regarded as a self-pollinating species, an exceptionally high heterozygosity observed in accessions, namely, *SB4* and *SB21* from the western geographic region, is an interesting result of this study, and should be further investigated.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

MG and ME designed the experiment and analyzed the data. ME conducted the experiment and wrote the draft manuscript. AC, CH, KT, MG, and TF reviewed the manuscript. All authors conceived the study and read and approved the submission of the manuscript for publication.

FUNDING

This research was financially supported by the Swedish International Development Cooperation Agency (Sida) and the Research and Training Grant awarded to the Addis Ababa University and the Swedish University of Agricultural Sciences (AAU-SLU Biotech; https://sida.aau.edu.et/index.php/biotechnology-phd-program/; accessed on September 25, 2021).

ACKNOWLEDGMENTS

We thank the Swedish International Development Cooperation Agency (Sida) for financing this research. We would also like to thank the Institute of Biotechnology, Addis Ababa University and Department of Plant Breeding, Swedish

University of Agricultural Sciences, for technical support during the course of the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.799482/full#supplementary-material

Supplementary Figure 1 | Geographical maps of Ethiopia showing **(A)** the original sampling locations of the sorghum accessions with red, green, and blue colors to highlight the western, northern, and eastern geographic regions, respectively and **(B)** the agro-ecological zones of Ethiopia as per the Global 16 Class classification system by Amede et al. (2015).

Supplementary Figure 2 | The panicle diversity of sorghum landraces grown in the country.

Supplementary Figure 3 | Minor allelic frequency (MAF) range of 3,001 SNP markers used for genetic diversity and population structure analyses of 359 individual plants representing the 24 sorghum accessions.

Supplementary Figure 4 | Principal coordinates analysis (PCoA) showing the clustering pattern of the 359 individuals of sorghum landraces and individuals denoted by the same color labels and shapes belonging to the same geographic region.

Supplementary Figure 5 | Inferred population structure of 24 sorghum accessions at K = 2. ΔK plot showing its maximum value at K = 2 suggesting two as the optimal number of genetic populations.

REFERENCES

Adugna, A. (2014). Analysis of in situ diversity and population structure in Ethiopian cultivated *Sorghum bicolor* (L.) landraces using phenotypic traits and SSR markers. *SpringerPlus* 3, 1–14. doi: 10.1186/2193-1801-3-212

Adugna, A., Snow, A. A., Sweeney, P. M., Bekele, E., and Mutegi, E. (2013). Population genetic structure of in situ wild *Sorghum bicolor* in its Ethiopian center of origin based on SSR markers. *Genet. Resour. Crop Evol.* 60, 1313–1328. doi: 10.1007/s10722-012-9921-8

Afolayan, G., Deshpande, S., Aladele, S., Kolawole, A., Angarawai, I., Nwosu, D., et al. (2019). Genetic diversity assessment of sorghum (*Sorghum bicolor* (L.) Moench) accessions using single nucleotide polymorphism markers. *Plant Genet. Resour.* 17, 412–420.

Ali, M., Rajewski, J., Baenziger, P., Gill, K., Eskridge, K., and Dweikat, I. (2008). Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. *Mol. Breed.* 21, 497–509. doi: 10.1007/s11032-007-9149-z

Amede, T., Auricht, C., Boffa, J.-M., Dixon, J. A., Mallawaarachchi, T., Rukuni, M., et al. (2015). *The evolving farming and pastoral landscapes in ethiopia: a farming system framework for investment planning and priority setting*. Canberra, ACT: ACIAR.

Ayana, A., and Bekele, E. (2000). Geographical patterns of morphological variation in sorghum (*Sorghum bicolor* (L.) Moench) germplasm from Ethiopia and Eritrea: quantitative characters. *Euphytica* 115, 91–104.

Ayana, A., Bryngelsson, T., and Bekele, E. (2000). Genetic variation of Ethiopian and Eritrean sorghum (*Sorghum bicolor* (L.) Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). *Genet. Resour. Crop Evol.* 47, 471–482. doi: 10.1111/j.1601-5223.2000.t01-1-00249.x

Barnaud, A., Deu, M., Garine, E., Mckey, D., and Joly, H. I. (2007). Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. *Theor. Appl. Genet.* 114, 237–248. doi: 10.1007/s00122-006-0426-8

Becker, C., Fischer, J., and Münitz, K. (1994). PCR cloning and expression analysis of cDNAs encoding cysteine proteinases from germinating seeds of Vicia sativa L. *Plant Mol. Biol.* 26, 1207–1212. doi: 10.1007/BF00040701

Borrell, A. K., Hammer, G. L., and Douglas, A. C. (2000). Does maintaining green leaf area in sorghum improve yield under drought? I. Leaf growth and senescence. *Crop Sci.* 40, 1026–1037.

Brondani, C., Caldeira, K. D. S., Borba, T. C. O., Pn, R., De Morais, O. P., Castro, E. D. M., et al. (2006). Genetic variability analysis of elite upland rice genotypes with SSR markers. *Embrapa Arroz Feijão Artigo periódico indexado* 6, 9–17. doi: 10.12702/1984-7033.v06n01a02

Bucheyeki, T. L., Gwanama, C., Mgonja, M., Chisi, M., Folkertsma, R., and Mutegi, R. (2009). Genetic variability characterisation of Tanzania sorghum landraces based on simple sequence repeats (SSRs) molecular and morphological markers. *Afr. Crop Sci. J.* 17:54201.

Burow, G., Franks, C. D., Xin, Z., and Burke, J. J. (2012). Genetic diversity in a collection of Chinese sorghum landraces assessed by microsatellites. *Am. J. Plant Sci.* 3, 1722–1729. doi: 10.4236/ajps.2012.312210

Cortés, A. J., and Blair, M. W. (2018). Genotyping by sequencing and genome–environment associations in wild common bean predict widespread divergent adaptation to drought. *Front. Plant Sci.* 9:128. doi: 10.3389/fpls.2018.00128

Cuevas, H. E., and Prom, L. K. (2013). Assessment of molecular diversity and population structure of the Ethiopian sorghum [*Sorghum bicolor* (L.) Moench] germplasm collection maintained by the USDA–ARS National Plant Germplasm System using SSR markers. *Genet. Resour. Crop Evol.* 60, 1817–1830. doi: 10.1007/s10722-013-9956-5

Cuevas, H. E., and Prom, L. K. (2020). Evaluation of genetic diversity, agronomic traits, and anthracnose resistance in the NPGS Sudan Sorghum Core collection. *BMC Genomics* 21:88. doi: 10.1186/s12864-020-6489-0

Cuevas, H. E., Rosa-Valentin, G., Hayes, C. M., Rooney, W. L., and Hoffmann, L. (2017). Genomic characterization of a core set of the USDA-NPGS Ethiopian sorghum germplasm collection: implications for germplasm conservation, evaluation, and utilization in crop improvement. *BMC Genomics* 18, 1–17. doi: 10.1186/s12864-016-3475-7

De Alencar Figueiredo, L. F., Sine, B., Chantereau, J., Mestres, C., Fliedel, G., Rami, J.-F., et al. (2010). Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI, Ae1, Wx and O2. *Theor. Appl. Genet.* 121, 1171–1185. doi: 10.1007/s00122-010-1380-z

De Oliveira Borba, T. C., Brondani, R. P. V., Rangel, P. H. N., and Brondani, C. (2009). Microsatellite marker-mediated analysis of the EMBRAPA Rice Core Collection genetic diversity. *Genetica* 137, 293–304. doi: 10.1007/s10709-009-9380-0

De Wet, J., and Harlan, J. (1971). The origin and domestication ofSorghum bicolor. *Econ. Bot.* 25, 128–135.

Desmae, H., Jordan, D. R., and Godwin, I. D. (2016). Geographic patterns of phenotypic diversity in sorghum (*Sorghum bicolor* (L.) Moench) landraces from North Eastern Ethiopia. *Afr. J. Agricult. Res.* 11, 3111–3122.

Djè, Y., Heuertz, M., Lefebvre, C., and Vekemans, X. (2000). Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. *Theor. Appl. Genet.* 100, 918–925. doi: 10.1007/s001220051371

Earl, D. A. (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

Ejeta, G. (2005). "Integrating biotechnology, breeding, and agronomy in the control of the parasitic weed Striga spp in sorghum," in *the wake of the double helix: from the green revolution to the gene revolution, Bologna Bologna*, eds R. Tuberosa, R. L. Phillips, and M. Gale (Bologna: Avenue Media), 239–251.

Enyew, M., Feyissa, T., Geleta, M., Tesfaye, K., Hammenhag, C., and Carlsson, A. S. (2021). Genotype by environment interaction, correlation, AMMI, GGE biplot and cluster analysis for grain yield and other agronomic traits in sorghum (*Sorghum bicolor* L. Moench). *PLoS One* 16:e0258211. doi: 10.1371/journal.pone.0258211

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, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

FAOSTAT (2019). *Food and Agriculture Organization of the United Nations*. Rome: FAO.

Faye, J. M., Maina, F., Hu, Z., Fonceka, D., Cisse, N., and Morris, G. P. (2019). Genomic signatures of adaptation to Sahelian and Soudanian climates in sorghum landraces of Senegal. *Ecol. Evolut.* 9, 6038–6051. doi: 10.1002/ece3.5187

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*. [Preprint].

Geleta, N., and Labuschagne, M. (2005). Qualitative traits variation in sorghum (*Sorghum bicolor* (L.) Moench) germplasm from, eastern highlands of Ethiopia. *Biodivers. Conserv.* 14, 3055–3064. doi: 10.1007/s10531-004-0315-x

Ghebru, B., Schmidt, R., and Bennetzen, J. (2002). Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. *Theor. Appl. Genet.* 105, 229–236. doi: 10.1007/s00122-002-0929-x

Girma, G., Nida, H., Seyoum, A., Mekonen, M., Nega, A., Lule, D., et al. (2019). A large-scale genome-wide association analyses of Ethiopian sorghum landrace collection reveal loci associated with important traits. *Front. Plant Sci.* 10:691. doi: 10.3389/fpls.2019.00691

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. doi: 10.1155/2015/431487

Hamrick, J. L. (1983). The distribution of genetic variation within and among natural plant populations. *Genet. Conserv.* 1983, 335–363.

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, 1179–1191. doi: 10.1111/1755-0998.12387

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat. Genet.* 42, 1027–1030. doi: 10.1038/ng.684

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Liu, K., and Muse, S. P. (2005). *New Genetic Data Analysis S oftware*. Massachusetts, MA: Whitehead Institute.

Luo, H., Zhao, W., Wang, Y., Xia, Y., Wu, X., Zhang, L., et al. (2016). SorGSD: a sorghum genome SNP database. *Biotechnol. Biofuels* 9, 1–9.

Rozas, J., Sánchez-Delbarrio, J.C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19, 2496–2497. doi: 10.1093/bioinformatics/btg359

Mace, E. S., Tai, S., Gilding, E. K., Li, Y., Prentis, P. J., Bian, L., et al. (2013). Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat. Commun.* 4, 1–9. doi: 10.1038/ncomms3320

Manzelli, M., Pileri, L., Lacerenza, N., Benedettelli, S., and Vecchio, V. (2007). Genetic diversity assessment in Somali sorghum (*Sorghum bicolor* (L.) Moench) accessions using microsatellite markers. *Biodivers. Conserv.* 16, 1715–1730. doi: 10.1007/s10531-006-9048-3

Markert, J. A., Champlin, D. M., Gutjahr-Gobell, R., Grear, J. S., Kuhn, A., Mcgreevy, T. J., et al. (2010). Population genetic diversity and fitness in multiple environments. *BMC Evol. Biol.* 10, 1–13. doi: 10.1186/1471-2148-10-205

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

Mehmood, S., Bashir, A., Ahmad, A., Akram, Z., Jabeen, N., and Gulfraz, M. (2008). Molecular characterization of regional *Sorghum bicolor* varieties from Pakistan. *Pak. J. Bot* 40, 2015–2021.

Melchiorre, M. G., Chiatti, C., Lamura, G., Torres-Gonzales, F., Stankunas, M., Lindert, J., et al. (2013). Social support, socio-economic status, health and abuse among older people in seven European countries. *PLoS One* 8:e54856. doi: 10.1371/journal.pone.0054856

Menamo, T., Kassahun, B., Borrell, A., Jordan, D., Tao, Y., Hunt, C., et al. (2021). Genetic diversity of Ethiopian sorghum reveals signatures of climatic adaptation. *Theor. Appl. Genet.* 134, 731–742. doi: 10.1007/s00122-020-03727-5

Mengistu, G., Shimelis, H., Laing, M., Lule, D., Assefa, E., and Mathew, I. (2020). Genetic diversity assessment of sorghum (*Sorghum bicolor* (L.) Moench) landraces using SNP markers. *S. Afr. J. Plant Soil* 37, 220–226. doi: 10.1080/02571862.2020.1736346

Mofokeng, A., Shimelis, H., Tongoona, P., and Laing, M. (2014). A genetic diversity analysis of South African sorghum genotypes using SSR markers. *S. Afr. J. Plant Soil* 31, 145–152. doi: 10.1080/02571862.2014.923051

Mohammadi, S. A., and Prasanna, B. (2003). Analysis of genetic diversity in crop plants—salient statistical tools and considerations. *Crop Sci.* 43, 1235–1248. doi: 10.1186/s12864-017-3922-0

Morris, G. P., Ramu, P., Deshpande, S. P., Hash, C. T., Shah, T., Upadhyaya, H. D., et al. (2013). Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci.* 110, 453–458. doi: 10.1073/pnas.1215985110

Motlhaodi, T., Geleta, M., Bryngelsson, T., Fatih, M., Chite, S., and Ortiz, R. (2014). Genetic diversity in'ex-situ'conserved sorghum accessions of Botswana as estimated by microsatellite markers. *Austral. J. Crop Sci.* 8, 35–43.

Motlhaodi, T., Geleta, M., Chite, S., Fatih, M., Ortiz, R., and Bryngelsson, T. (2017). Genetic diversity in sorghum [*Sorghum bicolor* (L.) Moench] germplasm from Southern Africa as revealed by microsatellite markers and agro-morphological traits. *Genet. Resour. Crop Evol.* 64, 599–610. doi: 10.1007/s10722-016-0388-x

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia university press.

Nei, M., and Takezaki, N. (1983). Estimation of genetic distances and phylogenetic trees from DNA analysis. *Proc. 5th World Cong. Genet. Appl. Livstock. Prod.* 21, 405–412.

Ng'uni, D., Geleta, M., and Bryngelsson, T. (2011). Genetic diversity in sorghum (*Sorghum bicolor* (L.) Moench) accessions of Zambia as revealed by simple sequence repeats (SSR). *Hereditas* 148, 52–62. doi: 10.1111/j.1601-5223.2011.02208.x

Ng'uni, D., Geleta, M., Hofvander, P., Fatih, M., and Bryngelsson, T. (2012). Comparative genetic diversity and nutritional quality variation among some important Southern African sorghum accessions ['*Sorghum bicolor*'(L.) Moench]. *Austral. J. Crop Sci.* 6, 56–64.

Nidumukkala, S., Tayi, L., Chittela, R. K., Vudem, D. R., and Khareedu, V. R. (2019). DEAD box helicases as promising molecular tools for engineering abiotic stress tolerance in plants. *Crit. Rev. Biotechnol.* 39, 395–407. doi: 10.1080/07388551.2019.1566204

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

Peakall, R., and Smouse, P. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research–an update. *Bioinformatics* 28:2537e2539. doi: 10.1093/bioinformatics/bts460

Peng, X., Zhao, Y., Cao, J., Zhang, W., Jiang, H., Li, X., et al. (2012). CCCH-type zinc finger family in maize: genome-wide identification, classification and expression profiling under abscisic acid and drought treatments. *PLoS One* 7:e40120. doi: 10.1371/journal.pone.0040120

Petit, R. J., El Mousadik, A., and Pons, O. (1998). Identifying populations for conservation on the basis of genetic markers. *Conserv. Biol.* 12, 844–855.

Pfeifer, B., Wittelsbürger, U., Ramos-Onsins, S. E., and Lercher, M. J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* 31, 1929–1936. doi: 10.1093/molbev/msu136

Poehlman, J., and Sleper, D. (1979). *Breeding field crops*. Amsterdam: Springer.

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

Radosavljević, I., Satovic, Z., and Liber, Z. (2015). Causes and consequences of contrasting genetic structure in sympatrically growing and closely related species. *AoB Plants* 7:lv106. doi: 10.1093/aobpla/plv106

Ramu, P., Billot, C., Rami, J.-F., Senthilvel, S., Upadhyaya, H., Reddy, L. A., et al. (2013). Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. *Theor. Appl. Genet.* 126, 2051–2064. doi: 10.1007/s00122-013-2117-6

Rao, S. A., Rao, K. P., Mengesha, M., and Reddy, V. G. (1996). Morphological diversity in sorghum germplasm from India. *Genet. Resour. Crop Evol.* 43, 559–567. doi: 10.1007/bf00138832

Rao, V. R., and Hodgkin, T. (2002). Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell Tiss. Org. Cult.* 68, 1–19.

Ruiz-Chután, J. A., Salava, J., Janovská, D., Žiarovská, J., Kalousová, M., and Fernández, E. (2019). Assessment of genetic diversity in Sorghum bicolor using RAPD markers. *Genetika* 51, 789–803.

Salem, K. F., and Sallam, A. (2016). Analysis of population structure and genetic diversity of Egyptian and exotic rice (*Oryza sativa* L.) genotypes. *C. R. Biol.* 339, 1–9. doi: 10.1016/j.crvi.2015.11.003

Shete, S., Tiwari, H., and Elston, R. C. (2000). On estimating the heterozygosity and polymorphism information content value. *Theor. Popul. Biol.* 57, 265–271. doi: 10.1006/tpbi.2000.1452

Showalter, A. M., Keppler, B. D., Liu, X., Lichtenberg, J., and Welch, L. R. (2016). Bioinformatic identification and analysis of hydroxyproline-rich glycoproteins in Populus trichocarpa. *BMC Plant Biol.* 16, 1–34. doi: 10.1186/s12870-016-0912-3

Silva, K. J. D., Pastina, M. M., Guimarães, C. T., Magalhães, J. V., Pimentel, L. D., Schaffert, R. E., et al. (2021). Genetic diversity and heterotic grouping of sorghum lines using SNP markers. *Sci. Agricola* 78:39.

Singh, R., and Axtell, J. D. (1973). High Lysine Mutant Gene (hl that Improves Protein Quality and Biological Value of Grain Sorghum 1. *Crop Sci.* 13, 535–539.

Statista (2020). *Sorghum production worldwide in 2019/2020, by leading country (in 1,000 metric tons)*. Hamburg: Statista.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595.

Tesfaye, E., Sime, M., Tirfesa, A., and Bekele, A. (2013). *Socio-economic assessment of moisture stress sorghum growing areas of Miesso and Kobo districts*. Addis Ababa: Ethiopian Institute of Agricultural Research.

Too, E. J., Onkware, A. O., Were, B. A. I, Gudu, S., Carlsson, A., and Geleta, M. (2018). Molecular markers associated with aluminium tolerance in *Sorghum bicolor*. *Hereditas* 155, 1–13. doi: 10.1186/s41065-018-0059-3

Tsehay, S., Ortiz, R., Johansson, E., Bekele, E., Tesfaye, K., Hammenhag, C., et al. (2020). New transcriptome-based SNP markers for Noug (*Guizotia abyssinica*) and their conversion to KASP markers for population genetics analyses. *Genes* 11:1373. doi: 10.3390/genes11111373

Valpuesta, V., Lange, N. E., Guerrero, C., and Reid, M. S. (1995). Up-regulation of a cysteine protease accompanies the ethylene-insensitive senescence of daylily (*Hemerocallis*) flowers. *Plant Mol. Biol.* 28, 575–582. doi: 10.1007/BF00020403

Wang, M. L., Zhu, C., Barkley, N. A., Chen, Z., Erpelding, J. E., Murray, S. C., et al. (2009). Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. *Theor. Appl. Genet.* 120, 13–23. doi: 10.1007/s00122-009-1155-6

Whitt, S. R., Wilson, L. M., Tenaillon, M. I., Gaut, B. S., and Buckler, E. S. (2002). Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci.* 99, 12959–12962. doi: 10.1073/pnas.202476999

Wilkinson, P. A., Winfield, M. O., Barker, G. L., Allen, A. M., Burridge, A., Coghill, J. A., et al. (2012). CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics* 13, 1–6. doi: 10.1186/1471-2105-13-219

Wondimu, Z., Dong, H., Paterson, A. H., Worku, W., and Bantte, K. (2021). Genetic diversity, population structure and selection signature in Ethiopian Sorghum (*Sorghum bicolor* L.[Moench]) germplasm. *bioRxiv*. [Preprint].

Wu, Y., Huang, Y., Tauer, C., and Porter, D. R. (2006). Genetic diversity of sorghum accessions resistant to greenbugs as assessed with AFLP markers. *Genome* 49, 143–149. doi: 10.1139/g05-095

Xing, H., Fu, X., Yang, C., Tang, X., Guo, L., Li, C., et al. (2018). Genome-wide investigation of pentatricopeptide repeat gene family in poplar and their expression analysis in response to biotic and abiotic stresses. *Sci. Rep.* 8, 1–9. doi: 10.1038/s41598-018-21269-1

Yan, S., Wang, L., Zhao, L., Wang, H., and Wang, D. (2018). Evaluation of genetic variation among sorghum varieties from southwest China via genome resequencing. *Plant Genome* 11:170098. doi: 10.3835/plantgenome2017.11.0098

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.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 06 January 2022
doi: 10.3389/fpls.2021.756182

The Genetic Diversity of Enset (*Ensete ventricosum*) Landraces Used in Traditional Medicine Is Similar to the Diversity Found in Non-medicinal Landraces

Gizachew Woldesenbet Nuraga[1,2], Tileye Feyissa[3], Kassahun Tesfaye[3,4], Manosh Kumar Biswas[1], Trude Schwarzacher[1], James S. Borrell[5], Paul Wilkin[5], Sebsebe Demissew[6], Zerihun Tadele[7] and J. S. (Pat) Heslop-Harrison[1]*

[1] Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom, [2] Department of Horticulture, Wolkite University, Wolkite, Ethiopia, [3] Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia, [4] Ethiopian Biotechnology Institute, Addis Ababa, Ethiopia, [5] Natural Capital and Plant Health Department, Royal Botanic Gardens, Kew, London, United Kingdom, [6] Department of Plant Biology and Biodiversity Management, Addis Ababa University, Addis Ababa, Ethiopia, [7] Institute of Plant Sciences, University of Bern, Bern, Switzerland

OPEN ACCESS

Edited by:
Andrés J. Cortés, Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

Reviewed by:
Asrat Asfaw, International Institute of Tropical Agriculture (IITA), Nigeria
Abush Abebe, International Institute of Tropical Agriculture (IITA), Nigeria

***Correspondence:**
Gizachew Woldesenbet Nuraga
bahrangw@gmail.com

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

Received: *16 September 2021*
Accepted: *08 December 2021*
Published: *06 January 2022*

Citation:
Nuraga GW, Feyissa T, Tesfaye K, Biswas MK, Schwarzacher T, Borrell JS, Wilkin P, Demissew S, Tadele Z and Heslop-Harrison JS (2022) The Genetic Diversity of Enset (Ensete ventricosum) Landraces Used in Traditional Medicine Is Similar to the Diversity Found in Non-medicinal Landraces. Front. Plant Sci. 12:756182. doi: 10.3389/fpls.2021.756182

Enset (*Ensete ventricosum*) is a multipurpose crop extensively cultivated in southern and southwestern Ethiopia for human food, animal feed, and fiber. It has immense contributions to the food security and rural livelihoods of 20 million people. Several distinct enset landraces are cultivated for their uses in traditional medicine. These landraces are vulnerable to various human-related activities and environmental constraints. The genetic diversity among the landraces is not verified to plan conservation strategy. Moreover, it is currently unknown whether medicinal landraces are genetically differentiated from other landraces. Here, we characterize the genetic diversity of medicinal enset landraces to support effective conservation and utilization of their diversity. We evaluated the genetic diversity of 51 enset landraces, of which 38 have reported medicinal value. A total of 38 alleles across the 15 simple sequence repeat (SSR) loci and a moderate level of genetic diversity (H_e = 0.47) were detected. Analysis of molecular variation (AMOVA) revealed that only 2.4% of the total genetic variation was contributed by variation among the medicinal and non-medicinal groups of landraces, with an F_{ST} of 0.024. A neighbor-joining tree showed four separate clusters with no correlation to the use-values of the landraces. Except for two, all "medicinal" landraces with distinct vernacular names were found to be genetically different, showing that vernacular names are a good indicator of genetic distinctiveness in these specific groups of landraces. The discriminant analysis of the principal components also confirmed the absence of distinct clustering between the two groups. We found that enset landraces were clustered irrespective of their use-value, showing no evidence for

genetic differentiation between the enset grown for 'medicinal' uses and non-medicinal landraces. This suggests that enset medicinal properties may be restricted to a more limited number of genotypes, might have resulted from the interaction of genotype with the environment or management practice, or partly misreported. The study provides baseline information that promotes further investigations in exploiting the medicinal value of these specific landraces.

Keywords: conservation, *Ensete ventricosum*, genetic diversity, landrace, SSR markers, traditional medicine

INTRODUCTION

Enset (*Ensete ventricosum*; also called Abyssinian banana) is a herbaceous, monocarpic perennial plant that grows from 4 to 10 m in height. It resembles and is closely related to bananas in the genus *Musa,* and these, together with the monotypic genus *Musella,* form the family Musaceae (Borrell et al., 2019). Enset is a regionally important crop, mainly cultivated for starchy human food, animal feed, and fiber. It contributes to the food security and rural livelihoods of a quarter of the population of Ethiopia (Yemataw et al., 2016). It is resilient to extreme environmental conditions, especially to drought (Tsegaye and Struik, 2002) and it is considered a priority crop in areas where the crop is grown as a staple food (Brandt et al., 1997).

The use of indigenous plant species to treat several ailments such as cancer, toothache, and stomach ache in different parts of Ethiopia has been frequently reported (Chekole, 2017; Megersa et al., 2019; Tesfaye et al., 2020). In addition to the extensive use of enset as human food and animal feed, some enset landraces play a well-known and important role in traditional medicine due to their use in repairing broken bones and fractures, assisting the removal of placental remains following birth or an abortion, and for treatment of liver disease (Terefe and Tabogie, 1989; Tsehaye and Kebebew, 2006; Olango et al., 2014). In the comparison of different medicinal plant species, *Ensete ventricosum* was ranked first by the local people for its medicinal use (Tefera and Kim, 2019). A compound that has anti-bacterial and anti-fungal activities extracted from *E. ventricosum* (Hölscher and Schneider, 1998) can be related to the traditional medicinal use of the plant by society. The free amino acid composition analysis of enset landraces indicates that high arginine content could be the other reason for their medicinal properties, as it is associated with collagen formation, tissue repair, and wound healing *via* proline, and it may also stimulate collagen synthesis as a precursor of nitric oxide (Tamrat et al., 2020a). However, experimental studies on different enset landraces claimed to have traditional medicinal importance are scant.

Enset production has been constrained by various plant pests, diseases, and abiotic factors (Merga et al., 2019; Kidane et al., 2021). The loss of some valuable enset genotypes due to various human and environmental factors was also previously reported (Gebremaryam, 1996; Negash et al., 2002). The existence of a gap in collections and conservation of enset landraces was also reported (Dalle and Daba, 2021). Medicinal landraces may be more threatened than others because when a person is ill, the medic is usually given the plant (free of charge) to cure the ailment of the patient, but the farmer does not have an economic reason to propagate and replant the medicinal landraces. Moreover, these landraces are highly preferred by wild animals like porcupines and wild pigs (Negash, 2007) and are more susceptible to diseases and drought (Nuraga et al., 2019a). Since these factors might lead to the complete loss of some of these important landraces, attention needs to be given to the conservation and proper utilization of the landraces that play important roles in traditional medicine.

Conserving domesticated enset diversity as seeds have been considered challenging for several reasons (Tamrat et al., 2020b), and the existing seed conservation measures of the enset crop and its wild relatives is insufficient (Guzzon and Müller, 2016). The most common method of conserving the genetic resources of vegetatively propagated plants like enset is in a field gene bank, which is very costly in terms of requirements for land, maintenance, and labor. In such cases, a clear understanding of the extent of genetic diversity is essential to reduce unnecessary duplication of germplasm (Rao and Hodgkin, 2002). Assessment of diversity using phenotypic traits is relatively straightforward and low cost (Cholastova and Knotova, 2012), and is the first step in identifying duplicates of accessions from phenotypically distinguishable cultivars. However, due to the influence of the environment on the phenotype, evaluating genetic variation at the molecular level is important.

Molecular markers are powerful tools in the assessment of genetic diversity which can assist the management of plant genetic resources (Virk et al., 2000; Teixeira da Silva et al., 2005). Previous enset genetic diversity studies have used molecular techniques including amplified fragment length polymorphism (AFLP; Negash et al., 2002), random amplification of polymorphic DNA (RAPD; Birmeta et al., 2002), inter simple sequence repeat (ISSR; Tobiaw and Bekele, 2011), and simple sequence repeat (SSR; Getachew et al., 2014; Biswas et al., 2020). SSR markers are highly polymorphic, co-dominant and the primer sequences are generally well conserved within and between related species (Karaagac et al., 2014). Recently, (Gerura et al., 2019) and (Olango et al., 2015) have reported the measurement of genetic diversity of enset using SSR markers. The previous studies were carried out on landraces from specific locations, and there was no identification and diversity study on enset landraces used for traditional medicine and other landraces. Therefore, the current study was conducted to investigate the extent of genetic diversity and the relationship that exists within and among enset landraces used in traditional medicine and those having other use-values.

MATERIALS AND METHODS

Plant Material and Genomic DNA Extraction

Thirty-eight cultivated and named *E. ventricosum* landraces which are used in the treatment of seven different human diseases or disorders were identified with the help of knowledgeable village elders from four locations (administrative zones/special district) consisting of nine districts or special districts of the Southern, Nations, Nationalities, and Peoples (SNNP) regional state of Ethiopia (**Figure 1**). For comparison, 13 enset landraces that have other non-medicinal use values (principally used for human food) were also sampled. To test the consistency of naming of landraces within each location, up to four duplicate samples (based on their availability) were collected from different sites. Since the landraces are not scientifically characterized, each individual was considered as a separate sample so that a total of 92 plant samples were collected (**Supplementary Table 1**). The samples were collected from individual farmers' fields, located at 18 Kebele (the lowest tier of civil administration unit) from across the enset distribution. Since different landraces may have been given the same vernacular name at different locations (Olango et al., 2015), landraces having identical names, but originated from different locations were labeled by including the first letter of names of location after a vernacular name of the second landrace. Healthy young cigar leaf (a recently emerged leaf still rolled as a cylinder) tissue samples were collected from individual plants from November to March 2017 and they were stored in zip-locked plastic bags containing silica gel and preserved until the extraction of genomic DNA. The dried leaf samples were ground and genomic DNA was isolated from 150 mg of each pulverized leaf sample following the modified cetyltrimethylammonium bromide (CTAB) extraction protocol (Borsch et al., 2003).

PCR Amplification and Electrophoresis

Twenty-one enset SSRs primer-pairs (14 from Olango et al., 2015, and 7 from Biswas et al., 2020) were initially screened for good amplification, polymorphism, and specificity to their target loci using 15 samples. This led to the selection of 15 primer pairs to genotype the landraces (**Table 1**).

A PCR amplification was carried out in a 20 μl reaction volume containing 1.5 μl (100 mM) template DNA, 11.5 μl molecular reagent water, W 4502 (Sigma, St. Louis, MO, United States), 0.75 μl dNTPs (10 mM) (Bio line, London), 2.5 μl Taq buffer (10× Thermopol reaction buffer), 1.25 μl $MgCl_2$ (50 mM), 1 μl forward and reverse primers (10 mM), and 0.5 μl (5 U/μl) BioTaq DNA polymerase (Bioline, London)

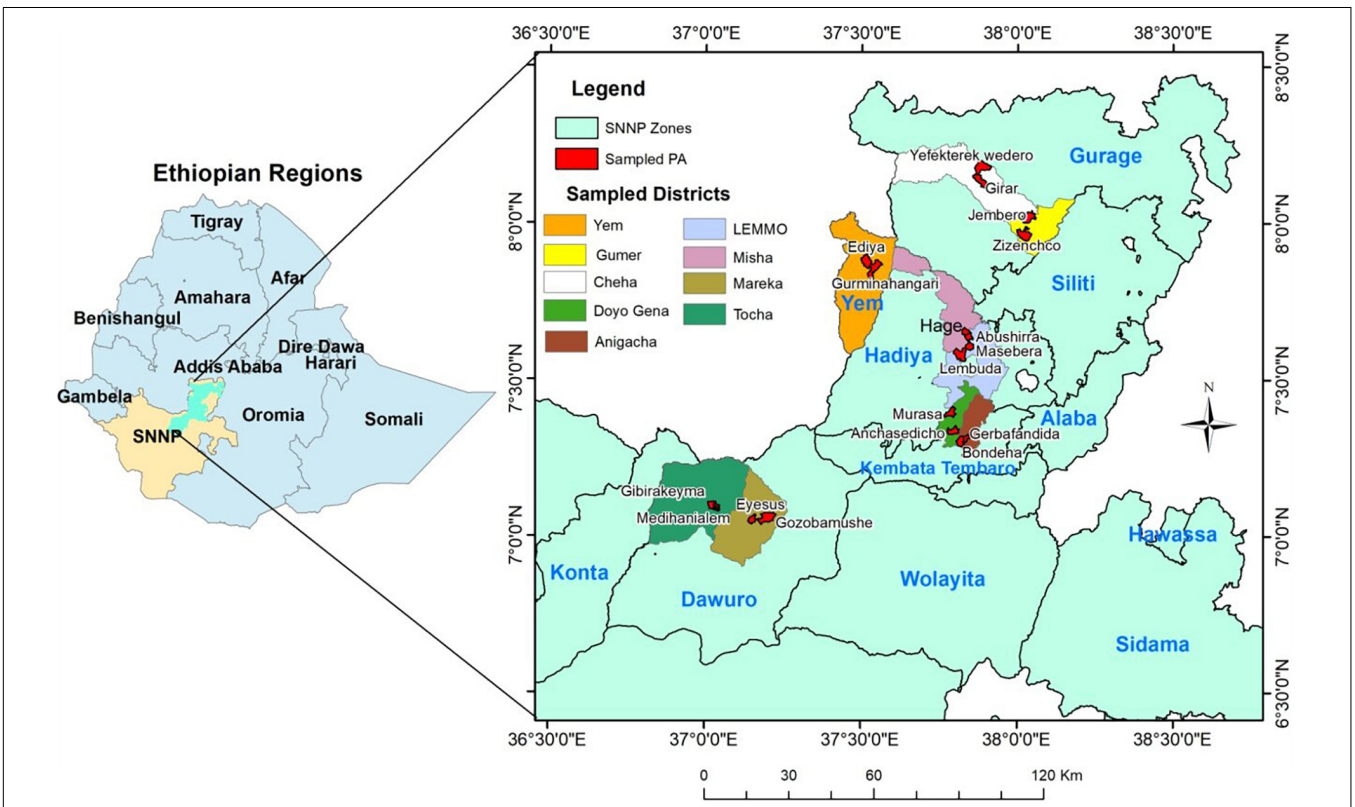


FIGURE 1 | Map of Ethiopia showing its Federal Regions **(left)** and enset sample collection sites that represent the nine studied districts found in, within four zones (Dawuro, Kembata-Tembaro, Hadya, and Gurage), and one special district (Yem) of the Southern, Nations, Nationalities, and Peoples (SNNP) Region. The map was constructed using geographic coordinates and elevation data collected from each sites using global positioning system (GPS). PA, Peasant association (the lowest tier of civil administration unit); SNNP, Southern, Nations, nationalities and Peoples.

TABLE 1 | Description and source of the 15 simple sequence repeat (SSR) primers used in genetic diversity of enset landraces.

Marker name	Forward primer sequence (5′–3′)	Reverse primer sequence (5′–3′)	Repeat motif	Size (bp)	References	T_a[a] (°C)
Evg-01	AGTCATTGTGCGCAGTTTCC	CGGAGGACTCCATGTGGATGAG	$(CTT)_8$	100–120	Olango et al., 2015	60
Evg-02	GGAGAAGCATTTGAAGGTTCTTG	TTCGCATTTATCCCTGGCAC	$(AG)_{12}$	118–153	Olango et al., 2015	62
Evg-04	GCCATCGAGAGCTAAGGGG	GGCAAGGCCGTAAGATCAAC	$(AG)_{21}$	113–147	Olango et al., 2015	60
Evg-05	AGTTGTCACCAATTGCACCG	CCATCCTCCACACATGCC	$(GA)_{22}$	103–141	Olango et al., 2015	62
Evg-06	CCGAAGTGCAACACCAGAG	TCGCTTTGCTCAACATCACC	$(GAA)_9$	202–211	Olango et al., 2015	62
Evg-08	CCATCGACGCCTTAACAGAG	TGAACCTCGGGAGTGACATAAG	$(GA)_{21}$	164–190	Olango et al., 2015	60
Evg-09	GCCTTTCGTATGCTTGGTGG	ACGTTGTTGCCGACATTCTG	$(GA)_{13}$	141–175	Olango et al., 2015	60
Evg-10	CAGCCTGTGCAGCTAATCAC	CAGCAGTTGCAGATCGTGTC	$(AG)_{21}$	191–210	Olango et al., 2015	60
Evg-11	GGCCTAGTGACATGATGGTG	TGATGCTAGATTCAAAGTCAAGG	$(AC)_{13}$	135–160	Olango et al., 2015	62
Evg-13	TTGAAAGCATTGCATGTGGC	TCACCACTGTAGACCTCAGC	$(CA)_{14}$	189–229	Olango et al., 2015	62
Evg-14	AACCAATCTGCCTGCATGTG	GCCAGTGATTGTTGAGGTGG	$(TGA)_8$	153–159	Olango et al., 2015	62
En[b]	ATCTGCATGCACCCTAGCTT	AAACCCTAACGTCCCTCCTC	$(GT)_{10}$	189	Biswas et al., 2020	62
En[c]	ATCAAGGTCATGTGCTGTGC	ATCAAGGTCATGTGCTGTGC	$(CT)_{11}$	116	Biswas et al., 2020	62
EnM00011571	GATCTGATCCACCTCCTCGT	CGACAAGGATCAAAATGGCT	$(AGG)_5$	277	Biswas et al., 2020	64
En[d]	TTCTCTTGCTGCACACACC	TCATGATCCCTGTCCTCCTC	$(GA)_9$	313	Biswas et al., 2020	64

T_a[a], annealing temperature; En[b], EnOnjSSR049028 marker; En[c], EnBedSSR020585 marker; En[d], EnM00025665 marker.

and amplified using a PCR thermal cycler (BiometraTOne, Terra Universal, Germany). The three-step amplification program consisted of initial (1) denaturation for 2 min at 95°C, (2) 35 cycles of denaturation at 95°C for 1 min, annealing at a temperature specific to each primer set (**Table 2**), for 1 min, extension at 72°C for 1min, and (3) final extension at 72°C for 10 min. The PCR products were stored at 4°C until electrophoresis.

The separation of the amplified product was accomplished in a 4% (w/v) agarose (Bioline, London) gel in 1% (w/v) Tris-acetate-EDTA (TAE) buffer containing ethidium bromide, and electrophoresed at 80 V for 3 h. A standard DNA ladder of 100 bp (Q step 2, Yorkshire Bioscience Ltd., United Kingdom) was loaded together with the samples to estimate molecular weight. The banding pattern was visualized using a gel documentation system (NuGenius, SYNGENE, Cambridge, United Kingdom) and the pictures were documented for scoring.

TABLE 2 | Levels of diversity indices of the SSR loci.

SSR Loci	N_a	N_e	I	H_o	H_e	uH_e	PIC	F
Evg1	3.00	2.27	0.89	0.39	0.54	0.56	0.48	0.30
Evg2	3.00	2.43	0.97	0.42	0.59	0.60	0.52	0.29
Evg4	3.00	2.29	0.92	0.63	0.56	0.57	0.49	−0.12
Evg5	2.00	1.82	0.64	0.54	0.45	0.46	0.36	−0.20
Evg6	2.00	1.41	0.43	0.00	0.27	0.28	0.26	1.00
Evg8	3.00	2.21	0.89	0.51	0.54	0.56	0.50	0.07
Evg9	3.00	2.27	0.93	0.44	0.55	0.56	0.49	0.20
Evg10	2.00	1.82	0.63	0.00	0.44	0.45	0.36	1.00
Evg11	3.00	1.87	0.74	0.41	0.46	0.47	0.43	0.08
Evg13	3.00	2.16	0.85	0.56	0.54	0.55	0.44	−0.06
Evg14	2.00	1.92	0.67	0.64	0.48	0.49	0.37	−0.34
EnO28	3.00	2.70	1.04	0.58	0.63	0.64	0.59	0.08
EnB85	2.00	1.23	0.31	0.21	0.18	0.18	0.16	−0.12
EnM71	2.00	1.91	0.67	0.51	0.48	0.49	0.36	−0.06
EnM65	2.00	1.60	0.56	0.41	0.38	0.39	0.31	−0.09
Mean	2.53	1.99	0.74	0.42	0.47	0.48	0.41	0.14

N_a, number of different alleles; N_e, number of effective alleles; I, Shannon's information index; H_o, observed heterozygosity; H_e, expected heterozygosity; uH_e, unbiased expected heterozygosity; PIC, polymorphic information content; F, fixation Index; EnO28, EnOnjSSR049028; EnB85, EnBedSSR020585; EnM71, EnM00011571; EnM65, EnM00025665.

Data Scoring and Analysis

The sizes of the clearly amplified fragments were estimated across all the sampled landraces. The number of different alleles (N_a), the effective number of alleles (N_e), Shannon's information index (I), observed heterozygosity (H_o), expected heterozygosity (H_e), un-biased expected heterozygosity (uH_e), and Fixation index for each locus were computed using GENALEX version 6.503 (Peakall and Smouse, 2012). The Polymorphism Information Content (PIC) for each locus was computed using PowerMarker version 3.25 (Liu and Muse, 2005). The Genetic differentiation (F_{ST}) between the two groups of landraces was estimated using GENALEX. An analysis of molecular variation (AMOVA) was performed to evaluate the relative level of genetic variations among groups, and among individuals within a group using GENALEX. The neighbor-joining (NJ) tree was constructed using the software DARwin (Perrier et al., 2003) based on Nei's genetic distance (Nei, 1972) to reveal the genetic relationships among the groups and individual landraces. The resulting trees were displayed using Fig Tree var.1.4.3 (Andrew, 2016). Discriminant Analysis of Principal Components (DAPC) was implemented using R, version 4.4.1 in 'adegenet' package (Jombart, 2008). Detection of admixture was inferred using a Bayesian model-based clustering algorithm implemented in STRUCTURE version 2.3.4 (Pritchard et al., 2000), To determine the most likely number of populations (K), the simulation method of Evanno et al. (2005) was implemented using the web-based STRUCTURE HARVESTER ver. 0.6.92 (Earl and von Holdt, 2012). Each of the probable K was run

10 times with K = 1–10, and the length of burning period was set at 50,000 and 500,000 Markov chain Monte Carlo (MCMC) iterations.

Principal coordinates analysis was carried out using R version 3.6.3 (R Core Team, 2017) to further evaluate the genetic similarity between the landraces.

RESULTS

Fifteen SSR markers that produced clear and scorable bands were analyzed to evaluate the genetic diversity and the relationship of *E. ventricosum* landraces used in traditional medicine and those having other use-values.

Genetic Diversity

The polymorphic nature of some of the SSR markers was as shown in **Supplementary Figure 1**. A total of 38 alleles were detected across 15 SSR loci in 92 genotypes (**Table 2**). The number of alleles generated per locus ranged from 2 to 3, with an average of 2.53 alleles. The PIC values for the markers varied from 0.16 (primer EnBedSSR020585) to 0.52 (primer Evg2) with an average of 0.41. The observed heterozygosity (H_o) and expected heterozygosity (H_e) ranged from 0 to 0.64 and 0.18 to 0.63, respectively, and Shannon's information index (I) ranged from 0.31 to 1.04.

Genetic Differentiation and Relationships

The AMOVA showed that 97.6% of the total variation was assigned to individuals within a group; while only 2.4% variation was contributed by variation among the groups (**Table 3**). The overall genetic divergences among the two groups of enset landraces ("medicinal" and "non-medicinal"), measured in coefficients of genetic differentiation (F_{ST}) was 0.024 (**Table 3**).

The unweighted neighbor-joining tree cluster analysis performed using Nei's genetic distance showed that landraces used in the treatment of a specific disease traditionally were not grouped into the same cluster or sub-cluster; instead, they were mixed with those landraces having other use values (**Figure 2**). Similarly, the landraces originating from each location were scattered into all 4 clusters (data not presented). In the neighbor-joining tree, it was also observed that some of the landraces with the same vernacular name (replicate samples) were found to be identical, while the others show a difference. Whereas, except two (*bishaeset* and *mekelwesa*), all landraces with the different vernacular names were distinct. The Bayesian clustering result showed the presence of four subpopulations, with some shared admixture memberships (**Figure 3A**), which is in agreement with the results of the neighbor-joining tree. The DAPC grouped the studied enset landraces into four clusters irrespective of their use value (**Figure 3B** and **Supplementary Table 2**). Although the medicinal and non-medicinal landraces were not separately clustered, the majority of the later were grouped in to cluster 3 and 4.

DISCUSSION

Genetic Diversity

We assessed the genetic diversity of 38 cultivated enset landraces used in traditional medicine and 13 landraces that have non-medicinal values. According to our results, a moderate level of genetic diversity (H_e = 0.47) was detected. A relatively higher H_e values (0.55 and 0.59) of enset were reported (Getachew et al., 2014) and (Olango et al., 2015; Gerura et al., 2019), respectively. The value of I (0.74) in the current study was also lower as compared with the earlier report (1.08) (Gerura et al., 2019). The variation of the result is probably due to the fact that our study was focused on a selected group of landraces, those used in traditional medicine. Lower genetic diversity estimates were reported earlier using ISSR (Tobiaw and Bekele, 2011) and RAPD (Birmeta et al., 2002) markers. However, comparisons of detailed diversity estimates from marker systems that have different properties and origins of variation do not allow useful conclusions (Powell et al., 1996; Hamza et al., 2013).

Genetic Differentiation and Relationships

The genetic differentiation between the landraces used in traditional medicine and those having other use-values was very low (0.024). Genetic differentiation values (0.037) among locations were reported on enset (Gerura et al., 2019), although the direct comparison of different populations is difficult. The AMOVA also showed that the proportion of genetic variation among the two groups of enset landrace was very much limited (2.4%), while the majority was contributed by variation among individuals.

From the landraces that have the same vernacular names (replicate samples), the majority were closely similar genetically and placed together in the neighbor-joining tree. This indicates that farmers have rich indigenous knowledge in identifying and naming enset landraces based on phenotypic traits, and the knowledge is shared across the growing region. However, few other replicates of landraces were placed in different clusters, indicating that genetically different landraces were given the same vernacular name. Perfect identification of genotypes using morphological traits is difficult, and the existence of homonyms has been reported previously (Olango et al., 2015).

TABLE 3 | Analysis of molecular variance and fixation index for landraces used in traditional medicine and those having other use values based on data from 15 loci.

Source of variation	Degree of freedom	Sum of square	Variance components	Percent variation	Fixation index	*P* value
Among groups	1	8.28	0.09	2.4	F_{ST}: 0.024	0.008
Within groups	182	669.83	3.68	97.6		
Total	183	678.11	3.77	100		

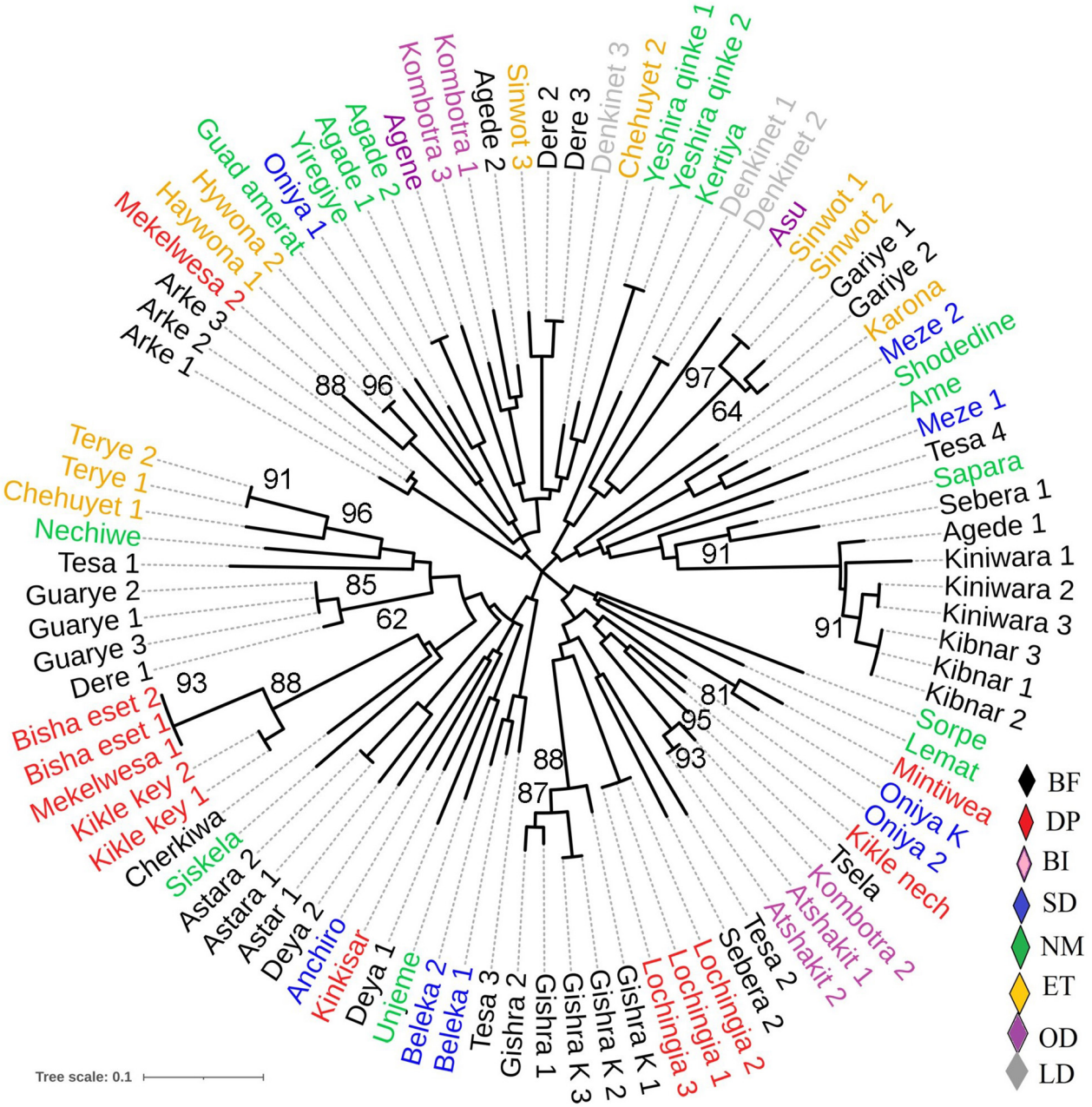


FIGURE 2 | The unrooted neighbor-joining tree generated based on simple matching dissimilarity coefficients over 1,000 replicates, showing the genetic relationship among 51 *Ensete ventricosum* landraces (duplicated on average two times) using 15 SSR markers. Landraces are color-coded according to previously identified diseases types or disorders treated by the landraces traditionally, as designated by: BF for bone fracture; DP, discharge of placenta; BI, back injury; SD, skin itching and diarrhea; NM, non-medicinal; ET, expulsion of thorn and drainage of abscess from tissue; LD, liver disease; OD, other diseases. Numbers at the nodes are bootstrap values only ≥ 60%.

Except for two, all landraces with distinct vernacular names were found to be different, showing that vernacular names are good indicators of genetic distinctiveness in these specific groups of landraces. Whereas, the existence of 37 and 8 duplicates of landrace in diversity analysis of enset using four AFLP (Negash et al., 2002) and 12 RAPD (Birmeta et al., 2002) markers, respectively, was reported. Gerura et al. (2019), who studied 83 enset genotypes using 12 SSR markers, also reported 10 duplicates of landraces. Although full identity among the landraces can only be determined when the entire genomes are compared, it is expected that the SSR markers used in the current study could sufficiently discriminate the landraces than the studies that reported a higher number of duplicates. The variation of the results, therefore, could be due to the sample collection method followed in the current study, which involved focusing mainly on specific landraces used in traditional medicine.

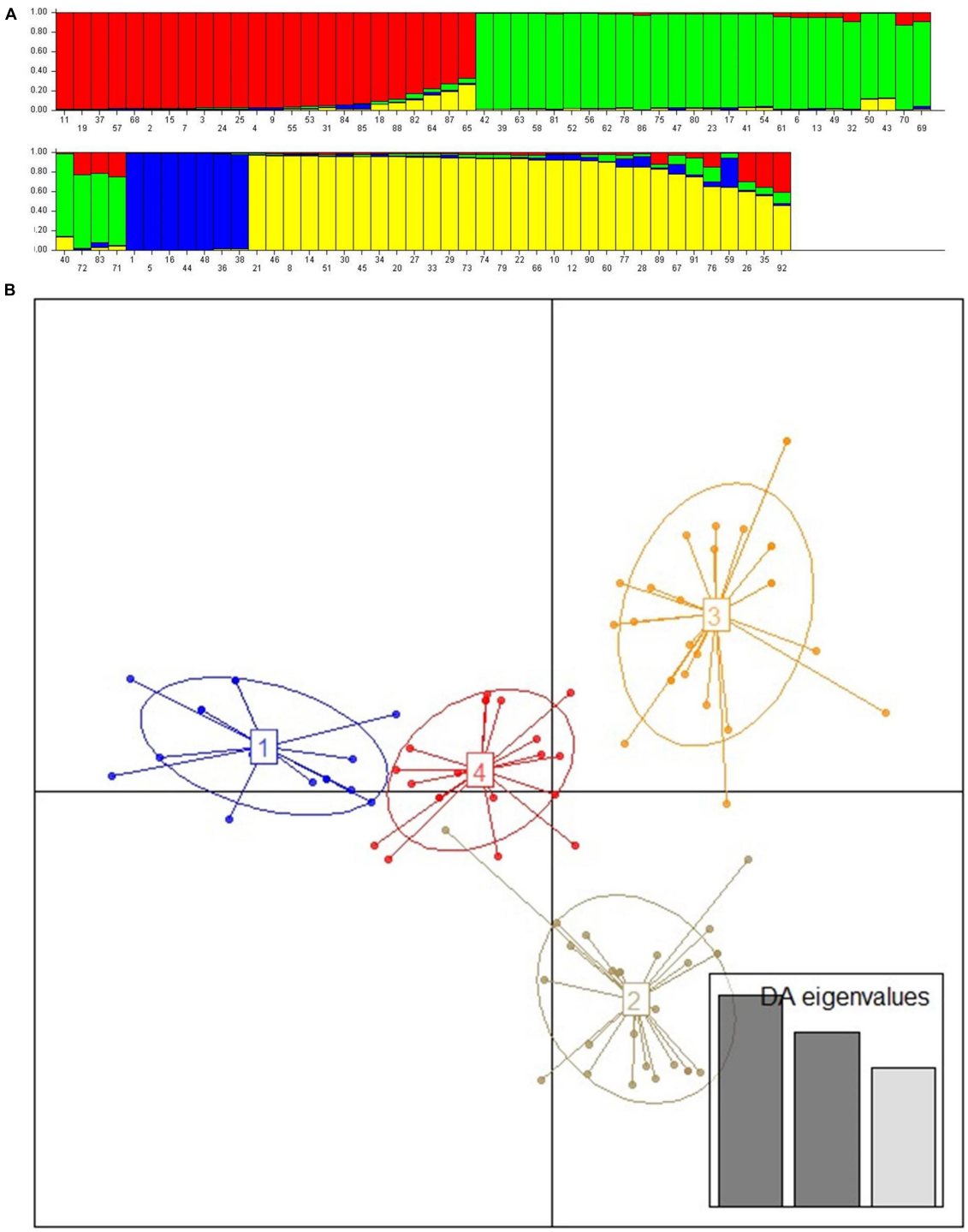


FIGURE 3 | Population structure and detection of admixture based on 15 polymorphic simple sequence repeat (SSR) markers indicating estimated group structure with individual group membership values (1–92 following arrangement of landraces in **Supplementary Table 1**) **(A)** and Discriminant analyses of principal components (DAPC) scatter plot for 92 enset landraces **(B)**. The axes represent the first two linear discriminants, each circle represents a cluster, and each dot represents an individual. Numbers represent the different subpopulations identified by DAPC analysis.

The use of some of the enset landraces in traditional treatment of various human ailments in the major enset growing region of Ethiopia, SNNPR, was reported by several authors (Tsehaye and Kebebew, 2006; Olango et al., 2014; Ayenew et al., 2016; Daba and Shigeta, 2016). However, landraces that are used in the treatment of the same types of diseases did not

show distinct grouping; instead, landraces used to treat different diseases were mixed with each other and even with those having other use values in the neighbor-joining tree, indicating that "medicinal" properties do not appear to be monophyletic. Furthermore, the DAPC also showed that the two groups of landraces neither formed a separate cluster nor did one group show greater spread or genetic diversity. From these results, it can be argued that landraces that are used in traditional medicine are not genetically distinct from other landraces.

There are several possible explanations for these observations. First, all enset landraces may have a degree of medicinal value, but specific genotypes are preferred for phenotypic or cultural reasons. Second, the medicinal value may arise through genotype-environment interactions or management practices specific to those landraces i.e., they may have non-differentiated genotypes, but *in situ*, they generate unique biochemistry with medicinal properties. Thirdly, a number of important medicinal landraces may have been omitted, or medicinal value incorrectly assigned to non-medicinal landraces. This could serve to hinder our analysis and make it more difficult to detect real genetic differentiation. This would also be an indication of a decline in the quality of indigenous knowledge. We also note that it is unlikely that the strong trust of society upon these landraces could not be developed after a very long period of use, and we have observed remarkable similar enset medicinal claims across a wide variety of distinct ethnic groups in multiple languages. Moreover, anti-bacterial and anti-fungal activities of a compound extracted from the unspecified *E. ventricosum* landrace (Hölscher and Schneider, 1998), and a report (Sreekutty and Mini, 2016) on the medicinal property of a related species, *Ensete superbum*, suggests that at least some of the enset landraces have real medicinal property. The higher mineral concentration (that has a relation with bone health) landraces used in traditional medicine was reported (Nuraga et al., 2019b). Finally, a biochemical survey of enset landraces (Tamrat et al., 2020a) detected high levels of arginine, compared to other amino acids, in three medicinal landraces (*Koshkowashiye*, *Astara*, and *Lochingiya*). Arginine is involved in collagen formation, tissue repair, and wound healing via proline, indicating a possible biochemical basis for the medicinal properties of some of the landraces.

CONCLUSION

The study indicated the existence of moderate level genetic diversity among enset landraces used in traditional medicine. The majority of the variation was contributed by variation among individuals, indicating low genetic differentiation among the groups. Except for two, all the landraces with distinct vernacular names were found to be genetically different. The landraces were not clustered based on their use-values, showing no evidence for genetic differentiation between landraces used in traditional medicine and those having other use-values, and the range of diversity in medicinal landraces was little different from that of landraces cultivated for food. In the future, we suggest a biochemical comparison of enset landraces growing in the same environmental and soil condition would complement our analysis, while genetic mapping and genome-wide association studies (GWAS) have the potential to identify genomic regions and genes associated with medicinal traits. The information from this study will be useful for the identification and conservation of enset landraces used in traditional medicine, and it can provide baseline information that promotes further investigations in exploiting the medicinal value of these specific landraces.

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

GN, TF, KT, and SD designed the experiment in the context of the funded project designed by PW, JH-H, and SD. GN carried out the sample collection, primer design, and laboratory work with MB. GN and MB conducted data mining and carried out the data analysis. GN, JB, TF, and JH-H were major contributors to interpreting the data. All authors contributed background to the design of the work and manuscript writing and revision and approved the final manuscript.

FUNDING

This work was financially supported by Addis Ababa University through the Thematic Research Project and GCRF Foundation Awards for Global Agricultural and Food Systems Research through "Modeling and genomics resources to enhance the exploitation of the sustainable and diverse Ethiopian starch crop enset and support livelihoods" project Reference BB/P02307X/1. The first body was funded for expenses related to the field data collection and DNA extraction, while the second body covered expenses related to PCR and other laboratory inputs, and also United Kingdom research visit costs of the corresponding author.

ACKNOWLEDGMENTS

The authors thank Hawi Niguse and Shiferaw Alemu for their assistance during DNA extraction. Fikadu Gadissa, Umer Abdi, and Muluken Birara are appreciated for their help in the data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.756182/full#supplementary-material

Supplementary Figure 1 | DNA fragments amplified in selected enset landraces by simple sequence repeat (SSR) primer; a. Evg2 (22 samples) and b. EnM00011571 (16 samples) resolved in agarose gel electrophoresis.

REFERENCES

Andrew, R. (2016). *FigTree: Tree Figure Drawing Tool*. Edinburgh: University of Edinburgh.

Ayenew, A., Mulatu, A., Lemma, B., and Girma, D. (2016). An ethnobotanical study of enset (*Ensete ventricosum* (Welw) Cheesman) in angacha woreda, kembata-tembaro zone, South Region, Ethiopia. *Am. J. Life Sci.* 4, 195–204. doi: 10.11648/j.ajls.20160406.18

Birmeta, G., Nybom, H., and Bekele, E. (2002). RAPD analysis of genetic diversity among clones of the Ethiopian crop plant *Ensete ventricosum*. *Euphytica* 124, 315–325.

Biswas, M. K., Darbar, J. N., Borell, J. S., Bagch, M., Biswas, D., Nuraga, G. W., et al. (2020). The landscape of microsatellites in the enset (*Ensete ventricosum*) genome and web_based marker resource development. *Sci. Rep.* 10:15312. doi: 10.1038/s41598-020-71984-x

Borrell, J. S., Biswas, M. K., Goodwin, M., Blomme, G., Schwarzacher, T., Heslop-Harrison, J. S., et al. (2019). Enset in Ethiopia: a poorly characterized but resilient starch staple. *Ann. Bot.* 123, 747–766. doi: 10.1093/aob/mcy214

Borsch, T., Hilu, K. W., Quandt, D., Wilde, V., Neinhuis, C., and Barthlott, W. (2003). Noncoding plastid trnT-trnF sequences reveal a well resolved pHylogeny of basal angiosperms. *J. Evol. Biol* 16, 558–576. doi: 10.1046/j.1420-9101.2003.00577.x

Brandt, S. A., Spring, A., Hiebsch, C., McCabe, J. T., Endale, T., Mulugeta, D., et al. (1997). *The "Tree Against Hunger": Enset- Based Agricultural Systems in Ethiopia*. Washington DC: American Association for the Advancement of Science.

Chekole, G. (2017). Ethnobotanical study of medicinal plants used against human ailments in Gubalafto District, Northern Ethiopia. *J. Ethnobiol. Ethnomed.* 13:55. doi: 10.1186/s13002-017-0182-7

Cholastova, T., and Knotova, D. (2012). Using morphological and microsatellite (SSR) markers to assess the genetic diversity in Alfalfa (*Medicago sativa* L.). *Int. J. Agric. Biosyst. Eng.* 6, 781–788.

Daba, T., and Shigeta, M. (2016). Enset (*Ensete ventricosum*) production in Ethiopia: its nutritional and socio-cultural values. *Agric. Food Sci. Res.* 3, 66–74.

Dalle, G., and Daba, D. (2021). Diversity and uses of enset [*Ensete ventricosum* (Welw.) Cheesman] varieties in Angacha district, Southern Ethiopia: call for taxonomic identifications and conservation. *Genet. Resour. Crop Evol.* 68:4. doi: 10.1007/s10722-020-00998-1

Earl, D. A., and von Holdt, B. M. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* 4, 359–361. doi: 10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14, 2611–2620. doi: 10.1111/j.1365-294X.2005.02553.x

Gebremaryam, S. (1996). *Enset Research in Ethiopia: 1976–1984. Enset Based Sustainable Agriculture in Ethiopia*. Addis Ababa: Institute of Agricultural Research.

Gerura, F. N., Meressa, B. H., Martina, K., Tesfaye, A., Olango, T. M., and Nasser, Y. (2019). Genetic diversity and population structure of enset (*Ensete ventricosum* Welw Cheesman) landraces of Gurage zone, Ethiopia. *Genet. Resour. Crop Evol.* 66, 1813–1824. doi: 10.1007/s10722-019-00825-2

Getachew, S., Mekibib, F., Admassu, B., Kelemu, S., Kidane, S., Negisho, K., et al. (2014). Look into genetic diversity of Enset (*Ensete ventricosum* (Welw.) Cheesman) using transferable microsatellite sequences of banana in Ethiopia. *J. Crop Improv.* 28, 159–183. doi: 10.1080/15427528.2013.861889

Guzzon, F., and Müller, J. V. (2016). Current availability of seed material of enset (*Ensete ventricosum*, Musaceae) and its Sub-Saharan wild relatives. *Genet. Resour. Crop Evol.* 63, 185–191.

Hamza, H., Abederrahim, M. A., Elbekkay, M., and Ferchichi, A. (2013). Comparison of the effectiveness of ISSR and SSR markers in determination of date palm (*Phoenix dactylifera* L.) agronomic traits. *Aust. J. Crop Sci.* 7, 763–769.

Hölscher, D., and Schneider, B. (1998). Phenylphenalenones from *Ensete ventricosum*. *Phytochemistry* 49, 2155–2157. doi: 10.1016/s0031-9422(98)00423-3

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Karaagac, E., Yilma, S., Cuesta-Marcos, A., and Vales, M. I. (2014). Molecular analysis of potatoes from the Pacific Northwest tri-state variety development program and selection of markers for practical DNA fingerprinting applications. *Am. J. Potato Res.* 91, 195–203. doi: 10.1007/s12230-013-9338-8

Kidane, S. A., Haukeland, S., Meressa, B. H., Hvoslef-Eide, A. K., and Coyne, D. L. (2021). Planting Material of Enset (*Ensete ventricosum*), a key food security crop in Southwest Ethiopia is a key element in the dissemination of plant-Parasitic nematode infection. *Front. Plant Sci.* 12:664155. doi: 10.3389/fpls.2021.664155

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

Megersa, M., Jima, T. T., and Goro, K. K. (2019). The use of medicinal plants for the treatment of toothache in Ethiopia. *Evid. Based Complement Alternat. Med.* 2019:2645174. doi: 10.1155/2019/2645174

Merga, I. F., Tripathi, L., Hvoslef-Eide, A. K., and Gebre, E. (2019). Application of genetic engineering for control of bacterial wilt disease of enset, Ethiopia's sustainability crop. *Front. Plant Sci.* 10:133. doi: 10.3389/fpls.2019.00133

Negash, A., Tsegaye, A., Van Treuren, R., and Visser, B. (2002). AFLP analysis of enset clonal diversity in south and Southwestern Ethiopia for conservation. *Crop Sci.* 42, 1105–1111.

Negash, F. (2007). *Diversity and Indigenous Management of Enset (Ensete ventricosum (Welw.) Cheesman) landraces in Guraghe Zone, Southern Ethiopia*. MSc thesis. Hawasa: Hawasa University.

Nei, M. (1972). Genetic distance between populations. *Amer. Naturalist* 106, 283–292.

Nuraga, G. W., Feyissa, T., Tesfaye, K., Demissew, S., and Tadele, Z. (2019a). Phenotypic diversity of enset (*Ensete ventricosum* (Welw.) Cheesman) landraces used in traditional medicine. *Genet. Resour. Crop Evol.* 66, 1761–1772.

Nuraga, G. W., Feyissa, T., Tesfaye, K., Demissew, S., and Zewdu, A. (2019b). Comparison of proximate, mineral and phytochemical composition of enset (*Ensete ventricosum* (Welw.) Cheesman) landraces used for a different purpose. *Afr. J. Agric. Res.* 14, 1326–1334.

Olango, T. M., Tesfaye, B., Catellani, M., and Pè, M. E. (2014). Indigenous knowledge, use and on-farm management of enset (*Ensete ventricosum* (Welw.) Cheesman) diversity in Wolaita, Southern Ethiopia. *J. Ethnobiol. Ethnomed.* 10:41. doi: 10.1186/1746-4269-10-41

Olango, T. M., Tesfaye, B., Pagnotta, M. A., Pè, M. E., and Catellani, M. (2015). Development of SSR markers and genetic diversity analysis in enset (*Ensete ventricosum* (Welw.) Cheesman), an orphan food security crop from Southern Ethiopia. *BMC Genet.* 16:98. doi: 10.1186/s12863-12015-10250-12868

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

Perrier, X., Flori, A., and Bonnot, F. (2003). "Data analysis methods," in *Genetic Diversity of Cultivated Tropical Plants*, eds P. Hamon, M. Seguin, X. Perrier, and J. C. Glaszmann (Montpellier: Enfield Science Publishers), 43–76.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., and Tingey, S. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Mol. Breed.* 2, 225–238.

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

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rao, V. R., and Hodgkin, T. (2002). Genetic diversity and conservation and utilization of plant genetic resources. *Plant Cell Tissue Organ Cult.* 68, 1–19.

Sreekutty, M. S., and Mini, S. (2016). Ensete superbum ameliorates renal dysfunction in experimental diabetes mellitus. *Iran J. Basic Med. Sci.* 19, 111–118.

Tamrat, S., Borrell, J. S., Biswas, M. K., Gashu, D., Wondimu, T., Vásquez-Londoño, C. A., et al. (2020a). Micronutrient composition and microbial community analysis across diverse landraces of the Ethiopian orphan crop enset. *Food Res. Int.* 137:109636. doi: 10.1016/j.foodres.2020.109636

Tamrat, S., Borrell, J. S., Shiferaw, E. K., Dickie, J. B., Nuraga, G. W., White, O., et al. (2020b). Germination ecology of wild and domesticated *Ensete ventricosum*: evidence for maintenance of sexual reproductive capacity in a vegetatively propagated perennial crop. *BioRxiv [preprint]* Available online at: https://doi.org/10.1101/2020.04.30.055582 (Accessed October 15, 2020).

Tefera, B. N., and Kim, Y. D. (2019). Ethnobotanical study of medicinal plants in the Hawassa Zuria District, Sidama zone, Southern Ethiopia. *J. Ethnobiol. Ethnomed.* 15:25. doi: 10.1186/s13002-019-0302-7

Teixeira da Silva, J. A., Nhut, D. T., Giang, D. T., and Rashid, S. Z. (2005). "Molecular markers for phylogeny, breeding and ecology in agriculture," in *Genetic Resources and Biotechnology*, Vol. III, eds D. Thangadurai, T. Pullaiah, and L. Tripathy (New Delhi: Regency Publications).

Terefe, B., and Tabogie, E. (1989). "A review of the available research recommendations and future strategies on enset," in *Proceedings of the National Crop Improvement Conference*, (Addis Ababa: Institute of Agricultural Research).

Tesfaye, S., Belete, A., Engidawork, E., Gedif, T., and Asres, K. (2020). Ethnobotanical study of medicinal plants used by traditional healers to treat cancer-like symptoms in eleven districts, Ethiopia. *Evid. Based Complement. Alternat. Med.* 2020:7683450. doi: 10.1155/2020/7683450

Tobiaw, D. C., and Bekele, E. (2011). Analysis of genetic diversity among cultivated enset (*Ensete ventricosum*) populations from Essera and Kefficho, southwestern part of Ethiopia using inter simple sequence repeats (ISSRs) marker. *Afr. J. Biotechnol.* 10, 15697–15709.

Tsegaye, A., and Struik, P. C. (2002). Analysis of enset (*Ensete ventricosum*) indigenous production methods and farm based biodiversity in major enset-growing regions of southern Ethiopia. *Expl. Agric.* 38, 291–315.

Tsehaye, Y., and Kebebew, F. (2006). Diversity and cultural use of enset (Enset ventricosum Welw.) Cheesman) in Bonga in situ conservation site, Ethiopia. *Ethnobotany Res. Appl.* 4, 147–157. doi: 10.17348/era.4.0.147-158

Virk, P. S., Newbury, J. H., Bryan, G. J., Jackson, M. T., and Ford-Lloyd, B. V. (2000). Are mapped or anonymous markers more useful for assessing genetic diversity? *Theor. Appl. Genet.* 100, 607–613. doi: 10.1007/s001220050080

Yemataw, Z., Tesfaye, K., Zeberga, A., and Blomme, G. (2016). Exploiting indigenous knowledge of subsistence farmers' for the management and conservation of Enset (*Ensete ventricosum* (Welw.) Cheesman) (musaceae family) diversity on-farm. *J. Ethnobiol. Ethnomed.* 12:34. doi: 10.1186/s13002-016-0109-8

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.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fpls.2022.837038

Leveraging National Germplasm Collections to Determine Significantly Associated Categorical Traits in Crops: Upland and Pima Cotton as a Case Study

Daniel Restrepo-Montoya[1], Amanda M. Hulse-Kemp[1,2], Jodi A. Scheffler[3], Candace H. Haigler[1,4], Lori L. Hinze[5], Janna Love[5], Richard G. Percy[5], Don C. Jones[6] and James Frelichowski[5]**

[1] Department of Crop and Soil Sciences, North Carolina State University, Raleigh, NC, United States, [2] Genomics and Bioinformatics Research Unit, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), Raleigh, NC, United States, [3] Crop Genetics Research Unit, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), Stoneville, MS, United States, [4] Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, United States, [5] Crop Germplasm Research Unit, United States Department of Agriculture - Agricultural Research Service (USDA-ARS), College Station, TX, United States, [6] Cotton Incorporated, Raleigh, NC, United States

OPEN ACCESS

Edited by:
Jinyoung Y. Barnaby, Agricultural Research Service (USDA), United States

Reviewed by:
Michael Benjamin Kantar, University of Hawai'i, United States
Bernardo Ordas, Spanish National Research Council (CSIC), Spain

***Correspondence:**
Amanda M. Hulse-Kemp
amanda.hulse-kemp@usda.gov
James Frelichowski
james.frelichowski@usda.gov

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

Received: 16 December 2021
Accepted: 21 March 2022
Published: 26 April 2022

Citation:
Restrepo-Montoya D, Hulse-Kemp AM, Scheffler JA, Haigler CH, Hinze LL, Love J, Percy RG, Jones DC and Frelichowski J (2022) Leveraging National Germplasm Collections to Determine Significantly Associated Categorical Traits in Crops: Upland and Pima Cotton as a Case Study. Front. Plant Sci. 13:837038. doi: 10.3389/fpls.2022.837038

Observable qualitative traits are relatively stable across environments and are commonly used to evaluate crop genetic diversity. Recently, molecular markers have largely superseded describing phenotypes in diversity surveys. However, qualitative descriptors are useful in cataloging germplasm collections and for describing new germplasm in patents, publications, and/or the Plant Variety Protection (PVP) system. This research focused on the comparative analysis of standardized cotton traits as represented within the National Cotton Germplasm Collection (NCGC). The cotton traits are named by 'descriptors' that have non-numerical sub-categories (descriptor states) reflecting the details of how each trait manifests or is absent in the plant. We statistically assessed selected accessions from three major groups of *Gossypium* as defined by the NCGC curator: (1) "Stoneville accessions (SA)," containing mainly Upland cotton (*Gossypium hirsutum*) cultivars; (2) "Texas accessions (TEX)," containing mainly *G. hirsutum* landraces; and (3) *Gossypium barbadense* (Gb), containing cultivars or landraces of Pima cotton (*Gossypium barbadense*). For 33 cotton descriptors we: (a) revealed distributions of character states for each descriptor within each group; (b) analyzed bivariate associations between paired descriptors; and (c) clustered accessions based on their descriptors. The fewest significant associations between descriptors occurred in the SA dataset, likely reflecting extensive breeding for cultivar development. In contrast, the TEX and Gb datasets showed a higher number of significant associations between descriptors, likely correlating with less impact from breeding efforts. Three significant bivariate associations were identified for all three groups, *bract nectaries:boll nectaries*, *leaf hair:stem hair*, and *lint color:seed fuzz color*. Unsupervised clustering analysis recapitulated the species labels for about

97% of the accessions. Unexpected clustering results indicated accessions that may benefit from potential further investigation. In the future, the significant associations between standardized descriptors can be used by curators to determine whether new exotic/unusual accessions most closely resemble Upland or Pima cotton. In addition, the study shows how existing descriptors for large germplasm datasets can be useful to inform downstream goals in breeding and research, such as identifying rare individuals with specific trait combinations and targeting breakdown of remaining trait associations through breeding, thus demonstrating the utility of the analytical methods employed in categorizing germplasm diversity within the collection.

Keywords: trait association, categorical data, cotton, crop germplasm, breeding

INTRODUCTION

Global agriculture production is facing major challenges, including demands to increase crop productivity and quality while sufficiently preserving natural ecosystems, addressing climate change and tolerance of intense weather events, increasing agricultural resource use efficiency, and enhancing biotic and abiotic stress resistance (FAO, 2017; Tian et al., 2021). To address these challenges, a constant interaction between plant breeding and fundamental research is needed, and both approaches have been used to address challenges of crop production for food, fiber, fuel, animal feeds, and ornamental uses, among others (Gillespie and van den Bold, 2017; Ramankutty et al., 2018; Zhao et al., 2019; Nguyen and Norton, 2020). Particularly, in the 21st century, agricultural intensification has relied on producing crops with genetic uniformity. Although these practices have benefits, they potentially increase crop susceptibility to pests, diseases, and environmental stress. To overcome those issues, the worldwide germplasm collections are essential to collecting and conserving living plant material, solving agricultural production problems, as well as conserving plant genetic diversity for future needs (Börner and Khlestkina, 2019; Nguyen and Norton, 2020). Among them, the largest collection in the world is the United States National Plant Germplasm System (NPGS) maintained by the United States Department of Agriculture - Agricultural Research Service (USDA-ARS). In the 1970's and 80's, the USDA mandated conservation of historical cultivars and crop wild relative germplasm for agricultural security (Wilkes and Williams, 2008). The NPGS is charged to acquire, conserve, document, distribute, evaluate, and characterize crop germplasm in order to safeguard the genetic diversity of agriculturally important plants (Allender, 2011; Byrne et al., 2018). Permanent collections and curators were established and available or acquired germplasm was re-routed to be first handled by the curators then maintained and distributed to users. There are currently 44 crop germplasm collections in the NPGS, the majority of which collect data on observable qualitative traits for each accession in the collections, including the National Cotton Germplasm Collection (NCGC) for *Gossypium* species (Postman et al., 2010; White et al., 2011).

Cotton is one of the most important cash crops around the world, and it provides the largest renewable source of fiber in addition to edible oil and protein (Campbell et al., 2010; Ahmad and Hasanuzzaman, 2020; Kumar et al., 2021). The NCGC began in 1989 and is physically maintained in College Station, TX, United States. It currently includes about 50 species of *Gossypium* and 10,459 total cotton accessions[1]. The collection is accompanied by information on the species classification and historical context of accessions, as traditionally described by a curator in the USDA-ARS Crop Germplasm Research Unit (CGRU). The NCGC primarily contains *G. hirsutum* and *G. barbadense*, which are the two main cultivated tetraploid cotton species (the other two cultivated types are diploids) (Grover et al., 2014). Upland cotton (*G. hirsutum* – Gh) and Pima cotton (*G. barbadense* – Gb), represent 75% of the total number of accessions in the NCGC collection. The Gh collection contains two main subsets as follows. (1) The Stoneville accessions (SA) mainly represent obsolete Gh cultivars originally collected at the Mississippi State University Delta Branch Experiment Station in Stoneville, Mississippi. (2) The Texas accessions (TEX) include photoperiodic landraces (i.e., primitive domesticated germplasm) or tropical materials as originally housed at Texas A&M University, College Station, Texas. The Pima accessions (Gb) were initially curated in Phoenix, Arizona, and the current group may contain a mix of landraces and cultivars, although specific subset information is not available (Percy et al., 2014).

In order to better characterize the diversity in the NCGC, a rating scale was established in 2006 for 36 phenotypic descriptors that encompass the diversity across *Gossypium* species in the collection, as observed by researchers in the CGRU (Yuan et al., 2021). For the past decade, the NCGC standardized and expanded descriptors to cover the consolidated sub-collection accessions and *Gossypium* species. However, the early stages of the cotton germplasm collection were sub-collections in different locations so historical descriptors and ratings differ. This systematic approach for describing traits has been used for evaluating many of the accessions in the NCGC over the last 11 years in the field in three different locations: (1) College Station, Texas; (2) Tecomán, Colima, Mexico; and (3) Liberia, Guanacaste, Costa Rica (Percy and Kohel, 1999; Wallace et al., 2008; Frelichowski and Percy, 2015; Yuan et al., 2021). Each of the 36 descriptors has a rating scale with a discrete number of non-numerical categories, or descriptor states, which encompass the

[1]https://npgsweb.ars-grin.gov/gringlobal/crop?id=547

variation in individual cotton accessions. Stated in another way, the rating scale for each trait contains a set number of categories or categorical variables, which may include for example presence, absence, and intermediary states of the trait between presence and absence (Percy et al., 2014; UPOV-Council, 2019; Cerda and Varoquaux, 2020).

Two of the cotton descriptors are illustrated in **Figure 1A**, leaf glands and leaf color. The rating scale for leaf glands has four descriptor states: glandless, light, medium, or heavy. The leaf glands descriptor is ordinal because there is a natural order within the range, but the distances between the states are not known. The rating scale for leaf color has three states: green, red, or dark red. The leaf color descriptor is nominal because its states are recognizable, but they lack inherent order. Neither nominal or ordinal variables have true quantitative values, but they can be evaluated through categorical analysis after grouping into a set of mutually exclusive unordered (nominal) or ordered (ordinal) categories (Watson, 2014; UPOV-Council, 2019) (**Figure 1B**). Classification of descriptor states into nominal and ordinal data types allows for the transformation of the data into a large matrix, and this, in turn, supports the use of statistical methods including bivariate association analysis to further characterize the large data set (**Figure 1C**). Bivariate association analysis determines whether or not there is a statistically significant relationship between any two descriptors within each group analyzed. Two descriptors are significantly associated if one of them tends to display specific states when the state of the other descriptor changes. Conversely, there is no significant association between two descriptors if their states change independently of each other (Watson, 2014; UPOV-Council, 2019). The evaluation and analysis of categorical traits have been previously suggested by the International Union for the Protection of New Varieties of Plants (UPOV-Council, 2019) as a means of demonstrating distinctness or statistically significant grouping patterns of different plant varieties.

Some examples of how categorical traits matter for cotton improvement are described below. The red color of cotton bolls, bracts, leaves, and stems may be useful for separating cotton genotypes during field tests, and it may also indicate enhanced resistance of red accessions to certain insects and/or pathogens (Long et al., 2019; Zhang et al., 2019). Likewise, the presence/absence of lysigenous glands, which contain terpenoid aldehydes including sesquiterpenoid gossypol, on bolls, leaves, and stems affects the degree of natural protection against insects. Conversely, the toxicity of these compounds to non-ruminant animals and humans limits the uses of cotton seeds and plant parts (Cai et al., 2010), which implies that breeders may want to alter the number and/or distribution of the glands (Zhou et al., 2013; Park et al., 2019; Gao et al., 2020). Other traits such as nectar glands on bolls, bracts, and leaves can, in an ecological context, provide nutrition for insects and microorganisms, while they promote insect damage in a crop context (Park et al., 2019). Moreover, the presence of hairs on leaves and stems may contribute to resistance to certain insects (i.e., Jassids) (Knight, 1952).

If meaningfully compared, the standardized phenotypic descriptors can be integrated with other phenotypic and genotypic data reported by NCGC to extract hidden information and expand the utility of the germplasm collection. We describe statistical methods to evaluate and extract additional meaning from phenotypic descriptors collected by a germplasm team. We leveraged a decade of collected data to compare descriptors within three major groups of *Gossypium* accessions maintained in the NCGC, including Pima cotton (Gb group) and cultivated Upland cotton (SA group) and its less-improved relatives (TEX group) (*G. hirsutum*). In this analysis, we add value to three of these sub-collections by identifying accessions that do have complete records for standardized phenotype descriptors and then exploring the descriptors. The results reveal: (a) distributions of character states for each descriptor within each of the three groups; (b) statistically significant bivariate associations between paired descriptors within each group; (c) label-blind, descriptor-based, clusters of accessions within a species; and (d) the ability to utilize clustering of descriptor data to identity the species of an accession. We anticipate that our prototypical analysis for cotton will be adaptable to the germplasm collections of other crops.

MATERIALS AND METHODS

Phenotypic Data and Accession Identification

A categorical analysis was applied using the phenotypic descriptors for selected accessions publicly available in the NCGC, which is part of the Germplasm Resource Information Network (GRIN)-Global (Cotton – see Text Footnote 1). The definition of the categorical descriptor scoring and methods for collecting the scores are reported in **Supplementary Data 1**. The scores correspond to standardized states for each descriptor that subdivide the overall range of the phenotype as observed in *Gossypium*. A particular descriptor may also include the "absent" state. The standardized descriptors and their rating scales are shown in table format on the CottonGen research community database website[2] (see also **Supplementary Table 1** – 05/01/2021), this is constantly updated as traits are added for evaluation.

The categorical descriptors for selected cotton accessions were obtained from the GRIN-Global system[3]. In the history of NCGC, a total of 28,258 observations on 10,459 accessions of 50 *Gossypium* species were made in the field between 1989 and 2019 (**Figure 2**). We studied data collected between 2011 and 2019 in correspondence with the time that observations began for 36 standardized descriptors under the direction of the USDA-ARS Crop Germplasm Research Unit (Wallace et al., 2008; Percy et al., 2014). In this last decade, a total of 11,616 observations (41% of the total set) on 7,941 unique accessions (as of May 2019) were in the database, but testing some of the accessions in multiple years and/or locations resulted in redundant records. Some of the records also had missing data points for one or more of the 36 descriptors. In order to obtain non-redundant

[2]https://www.cottongen.org/data/trait/NCGC_rating_scale

[3]https://www.grin-global.org

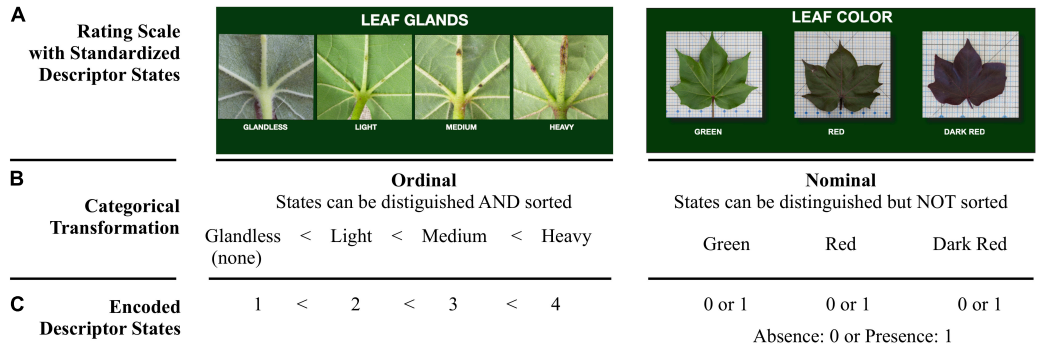


FIGURE 1 | Example of a categorical classification and encoding transformation of phenotypic descriptors. **(A)** Rating scales showing four (for leaf glands) or three (for leaf color) standardized descriptor states. **(B)** Examples of classifying two types of descriptor states, ordinal and nominal. **(C)** Encoding data transformation from text format to integer format prior to applying statistical analysis.

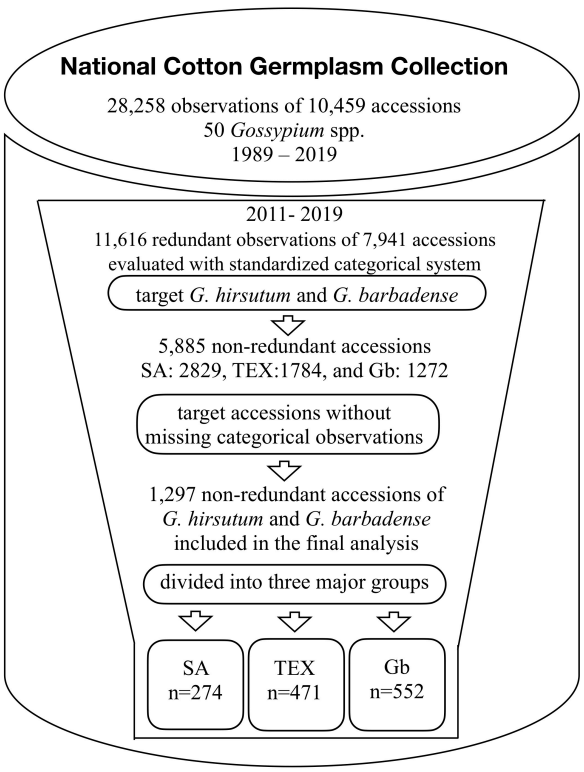


FIGURE 2 | Summary of the workflow leading to the final categorical trait descriptor dataset. Redundant observations reflect multiple evaluations of a single accession in multiple locations and/or different years. The data collected from 2011 to 2019 contained 11,616 field observations of approximately 21 *Gossypium* spp. Our work focused on data subsets enriched in *G. hirsutum* (Gh) and *G. barbadense* (Gb) accessions: SA, mainly Upland cotton cultivars; TEX, mainly photoperiodic/tropical landraces; and Gb – mainly Pima cotton. Numbers presented are as of May 2019, data is added to the database as it is collected so numbers are always increasing.

and complete records, the dataset was filtered using the following criteria. (1) Only accessions that belong to the SA, TEX, and Gb groups were selected. (2) The accessions with redundant records were randomly processed to select only one observation set per accession. (3) The accessions with missing information for any of the 36 descriptors were removed. After this filtering process, a total of 1,297 accessions with complete records were identified (SA, 274; TEX, 471; and Gb, 552, **Figure 2**). The accession IDs, the number of total seed requests per accession since 2007, and the associated descriptor information can be found in **Supplementary Tables 2–4**, for the SA, TEX, and Gb groups, respectively. The analysis finally included 46,692 data points.

Phenotypic Distributions and Data Transformations

For further analysis, 33 of 36 descriptors were retained because they were expected to be independent of the environment. Specifically, the scores for maturity, photoperiodic rating, and productivity were removed. The number of accessions in each group displaying each state of the analyzed descriptors was determined and displayed in distribution plots showing the observed variation across groups (**Supplementary Figure 1**). For statistical purposes, the descriptors were classified as nominal or ordinal prior to performing data transformations on the categorical scoring data of the remaining 33 descriptors (**Figure 1C** and **Supplementary Table 5**). All notations of segregation (seg)/off-type (i.e., where an accession was found to have varying levels of a descriptor) were removed from the analysis because the related phenotype was too complex or diverse to fit into the standardized rating scale for that descriptor (**Supplementary Data 1**). Only descriptors with two or more states observed in the field could be included in statistical analysis. To generate reasonable statistical power, each descriptor state was required to be represented by 5 or more accessions within the final data matrix. According to standard practice (Cochran, 1954; Camilli and Hopkins, 1978), some of the descriptor states were removed or combined if two or more of them together would include at least 5 observations (**Supplementary Table 6**). The changed instances were less than 5% of the initial data set that was used to plot phenotype distributions. This procedure explains why some descriptor states in the distribution plots are not also seen in the mosaic plots.

Bivariate Association and Contingency Analysis

The encoding transformation of the 33 descriptors produced a final data matrix for each group (SA, TEX, and Gb), which was then used for bivariate association analysis in JMP Pro 15.2.0 software (SAS Institute Inc., Cary, NC, United States). A contingency table was generated based on the comparison of each possible pair of descriptors. These tables show the number of observations for all of the different combinations of states of each descriptor. The contingency tables reveal how the states of descriptor 1 are contingent on the states of descriptor 2. We chose alpha = 0.01 as the standard for assessing significance. *P*-values were calculated by either Fisher's exact test (if both descriptors had only two states) or the Chi-square independent test (if at least one of the descriptors being compared had more than two states). The initial *p*-values were obtained as a list where each value corresponded to an independent bivariate association. Then the list was converted into a square matrix prior to adjusting for the False Discovery Rate (FDR) (Benjamini and Hochberg, 1995; Chiu, 2002). The FDR was only applied to the lower triangular matrix in order to avoid double-counting of the same comparison. The resulting FDR-corrected p-values of the bivariate associations were visualized in a heatmap for each of the three major groups of accessions. For each position in the heatmap, an associated mosaic plot shows a graphical representation of the two-way frequency table produced by the contingency analysis[4].

Unsupervised Analysis: Clustering Analysis and Multivariate Procedure

The same data set used for bivariate association analysis was used for unsupervised clustering analysis. K-modes clustering was used to explore similarities and/or differences among the three groups (see **Supplementary Tables 2–4**). The data matrix inclusive of all three groups was transformed using the scikit-learn 0.24.2 software (Pedregosa et al., 2011) into levels reflecting the rating scales of each descriptor prior to clustering analysis. K-modes unsupervised clustering analysis was run using kmodes version 0.11.0 (de Vos, 2021). The clustering analysis assumes a fixed number of clusters and tries to maximize the homogeneity within the clusters, so the analysis was run with $k = 2$ (aiming to discriminate Gh and Gb accessions) and $k = 3$ (aiming to discriminate SA, TEX, and Gb groups). The analysis depended on the prior encoding of the descriptors as nominal/ordinal, and the clustering was blind to NCGC labels for accession species/groups. The correlation of results of the algorithm species/group placement with the NCGC species/group labels was evaluated. Results were also evaluated by calculating an accuracy score of clustering using silhouette scores with the scikit-learn 0.24.2 software. The script implemented for this analysis is reported in **Supplementary Data 2** and the input file to run this analysis is reported in **Supplementary Table 7**.

An unsupervised multivariate procedure known as Multiple Correspondence Analysis (MCA) was also used to explore the relationship of the SA, TEX, and Gb accessions (Abdi and Valentin, 2007). In this procedure dimensionality reduction is applied over the categorical descriptors then identification of the non-linear interactions is performed. Afterward the first components are used to visualize the MCA "cloud of individuals" or the similarity structure of the accessions (see **Supplementary Tables 2–4**) (Kassambara, 2016; Nguyen and Holmes, 2019). The MCA analysis was applied in R, using the library FactoMineR version 2.4 (Lê et al., 2008) and the visualization was obtained using the library factoextra version 1.0.7 (Kassambra, 2022). The script implemented for this analysis is reported in **Supplementary Data 3**.

Bivariate Association Analysis Using Unsupervised Clustering Result

Three out of the nine sets identified by the unsupervised clustering analysis were reanalyzed using the bivariate association approach. The TEX accessions which clustered as TEX ($n = 308$), the SA accessions which clustered as TEX ($n = 156$) and the SA accessions which clustered as SA ($n = 251$) were processed (**Supplementary Table 9**). The remaining sets were not evaluated due to the low number of accessions clustered except for Gb, which largely contained the same set of samples as the prior bivariate association analysis.

Multiple Correspondence Analysis to Extract Information Content of Descriptors

The categorical traits were evaluated with the same MCA strategy as above (di Franco, 2016) to identify the contribution and correlation between descriptors. The contribution of each descriptor identified how much influence each categorical trait had in determining the overall information content relative to the entire set of traits (di Franco, 2016). The relationship between each of the variables was represented by calculating the correlation ratios between the accession coordinates on one component and each of the categorical variables, these results were visualized as the MCA "cloud of variables," or the similarity structure of categorical traits (Husson et al., 2010).

RESULTS

Cotton Accessions and Distributions of Descriptor States

The 33 cotton descriptors analyzed represent attributes evaluating vegetative, reproductive, and architectural structures of the plant for the three groups of accessions (SA, TEX, and Gb). Features such as color, nectaries, shape, or glands may be defined for multiple parts or aspects of the plant, with the different occurrences then counted as separate descriptors (**Table 1**). The three cotton groups analyzed often showed different patterns of variation for the states of each descriptor. There were cases where descriptor states were uniform in one group, but showed diverse distribution in others, and instances where each group displayed a different range of states for a particular descriptor.

[4]https://usda-ars-gbru.github.io/categorical_analysis_cotton/

TABLE 1 | Summary of 33 phenotypic descriptors analyzed in this study. Each descriptor, as marked by x, reflects a combination of a feature and the plant structure where the feature was evaluated.

		Plant structure													
		Boll	**Bract**	**Canopy**	**Fruiting**	**Growth**	**Leaf**	**Lint**	**Petal**	**Pollen**	**Seed fuzz**	**Seed**	**Stem**	**Locule**	**Stigma**
Feature	Color	x	x				x	x	x	x	x		x		
	Nectaries	x	x				x								
	Shape	x					x								
	Type		x	x	x							x			x
	Habit					x									
	Glands	x					x						x		
	Pitting	x													
	Pointing	x													
	Size	x					x								
	Teeth number		x												
	Teeth size		x												
	Hairs						x						x		
	Number													x	
	Spot								x						
	Density										x				

*Supporting document including the descriptor definitions (**Supplementary Data 1**).*

For example, glands are distributed across multiple parts of the plant and are, therefore, evaluated in bolls, leaves, and stems. The distributions within the different tissues showed that most SA and TEX accessions are medium or heavy glanded, whereas the Gb accessions were almost uniformly heavy glanded across all parts of the plant (**Figure 3**). Distributions for all descriptors in the three cotton groups analyzed are reported in **Supplementary Figure 1**.

Bivariate Associations of the Phenotypic Descriptors in Stoneville Accessions, Texas Accessions, and *Gossypium barbadense*

While plotting distributions of the states of individual descriptors across groups can be informative, it is also useful to identify cases where significant associations between descriptors occur within a group through bivariate association analysis. As an example, a breeder could ask the question: do the descriptor states of leaf glands change in parallel with the descriptor states of leaf hairs in different groups of *Gossypium* accessions? **Figure 4** shows heat maps displaying the significant *'descriptor_1:descriptor_2'* associations for the SA, TEX, and Gb groups independently. They show that the *'leaf glands:leaf hair'* comparison is significant for SA and TEX ($p \leq 0.01$). As previously mentioned, the association could not be analyzed in Gb because all of the accessions were heavy-glanded. Therefore, the 'leaf glands' descriptor does not appear in the Gb heat map.

From an overall perspective, the SA group had 23 significant associations out of 406 tested (all possible pairwise comparisons). The 23/406 ratio for SA (5.6%) compares to 153/406 for TEX (37.6%) and 122/351 for Gb (34.7%) (**Figure 4**). Among the three groups evaluated, most of the categorical descriptors show at least one significant association with another descriptor. The SA group had the largest number (9) of categorical descriptors with no significant association, meaning that its states changed independently of any other descriptor (stigma, seed fuzz, pollen color, locule number, leaf nectaries, growth habit, bract type, boll size, and boll point). Comparatively, all descriptors in TEX were significantly associated with at least one other descriptor, and Gb was similar with only one descriptor (boll point) lacking at least one association (**Figure 4**).

Examples of the contingency analysis are presented as mosaic plots, or stacked bar charts (**Figure 5**), which facilitate visual comparison of results between the groups analyzed. This type of plot was possible in cases where the descriptor had more than one state reported within the rating scale. These plots are important to analyze in cases of two or more of the cotton groups having the same significant *'descriptor_1:descriptor_2'* association, because the co-varying descriptor states may or may not be the same between groups (as shown here between **Figures 5A,B**). In each mosaic plot, the horizontal (*X*-) axis shows the states of descriptor_1 that were present in the group, with the width of each corresponding column portraying the proportion of accessions observed with that state of descriptor_1. The double vertical (*Y*-) axes together (black and blue arrows) describe descriptor_2, the vertical length of the bars is proportional to the number of accessions with each state of descriptor_2. The left-side *Y*-axis pertains to the proportion of descriptor_2 states found within the *X*-axis descriptor_1 variable states providing the overall likelihood that a trait state will be observed with the *X*-axis descriptor_1 trait state. The right-side *Y*-axis outlines the overall proportions of descriptor_2 (green arrow[5]). **Figure 5** shows the *'leaf glands:boll glanding'* mosaic plots for SA and TEX. In the SA group, most accessions had glands on bolls and

[5]https://www.jmp.com/en_us/statistics-knowledge-portal/exploratory-data-analysis/mosaic-plot.html

leaves, and a medium state of glanding dominated in both organs. Among rare accessions with glandless leaves, about 80% also had glandless bolls. On the contrary, the TEX group contained numerous accessions with heavy glanding in leaves and bolls, and no glandless associations were present using the baseline criteria of this study (**Figure 5**).

Relationships Between the Significant Descriptor Associations Existing in Stoneville Accessions, Texas Accessions, and *Gossypium barbadense*

The significant descriptor associations within each separate group (**Figure 4**) were intersected to identify commonalities and differences between the three groups, when possible, as shown in the Venn diagram (**Figure 6**). Most '*descriptor_1:descriptor_2*' evaluations were performed in all three groups (**Figure 6A**), but some descriptors were not analyzed in this way because they had the same state (homogeneous) in more than 98% of the accessions of one or more groups. These predominant homogeneous state phenotypes in each group were: for SA, leaf size (medium), seed type (free), bract teeth size (large), and bract teeth number (medium); for TEX, leaf color (green), leaf size (medium), seed type (free), and bract type (normal); and for Gb, leaf color (green), leaf shape (normal), stem glands (heavy), leaf glands (heavy), bract type (normal), and bract teeth number (medium) (**Supplementary Figure 1**). Of these predominant phenotypes, none are shared across all three groups, but three pairs are shared across two groups: SA and Gb both have medium bract teeth number and TEX and SA both have medium leaf size and the free seed type. In some other cases, a descriptor lacked multiple states in all three groups, which implied that only one or two groups could be compared. Correspondingly, the diagram in **Figure 6** is divided into sections showing comparisons between all three groups (SA vs. TEX vs. Gb, **Figure 6A**), across two groups (SA vs. TEX, TEX vs. Gb, or SA vs. Gb, **Figures 6B–D**) or only in one group (**Figures 6E–G**).

Across the three-group comparison, there were only three shared associations: '*bract nectaries:boll nectaries*', '*leaf hair:stem hair*', and '*lint color:seed fuzz color*' (**Figure 6A**). Breeders are concerned about nectaries due to their role in attracting insects, which often act as pests during production of cotton, given its capacity for self-pollination (Rudgers et al., 2004; Frelichowski and Percy, 2015; Zeng et al., 2018; Park et al., 2021). Here we use the names for nectary states as shown in **Figure 7A**. Both types of nectaries were 'present' in the majority of accessions analyzed for all three groups (**Figure 7B**). However, analysis of the mosaic plots shows some differences between the associated states of each descriptor between groups (**Figure 7C**). In the SA group, about 80% of the accessions had 'present' bract nectaries. Of these, about 80% also had boll nectaries. For the minority of accessions with reduced bract nectaries, about 95% of them also had reduced boll nectaries. Among the rare SA accessions that lacked bract nectaries, about 60% of them also lacked boll nectaries. However, for TEX, about 70% of the accessions had 'present' bract and boll nectaries. Of the remaining 30% with reduced bract nectaries, the boll nectaries were either 'present' or reduced in an approximately 50:50 ratio. Finally, the nectary traits in Gb were most similar to SA, but 'present' bract and boll nectaries existed in 99% of the accessions. When Gb bract nectaries were reduced in rare accessions, boll nectaries were either 'present' or reduced in an approximately 50:50 ratio (**Figure 7C**). The other two pairs of descriptor associations that were consistently found among the three accession groups ('*leaf hair:stem hair*', and '*lint color:seed fuzz color*') are further illustrated in **Supplementary Figures 4**, **5**, respectively.

Other bivariate descriptor associations were shared between only two groups or found in only one group. Between the SA and the TEX groups, six consistent associations were identified,

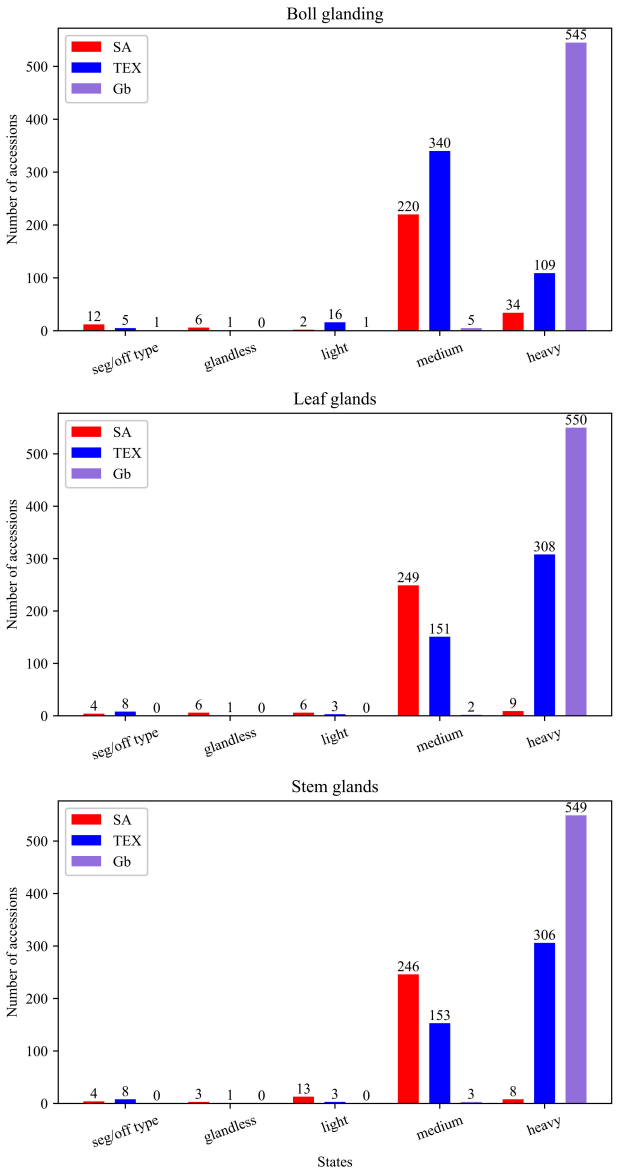


FIGURE 3 | Distribution of gland-related descriptors across plant structures and species. Seg/off type: segregation-off type condition, in which an accession was not homogeneous for a single state.

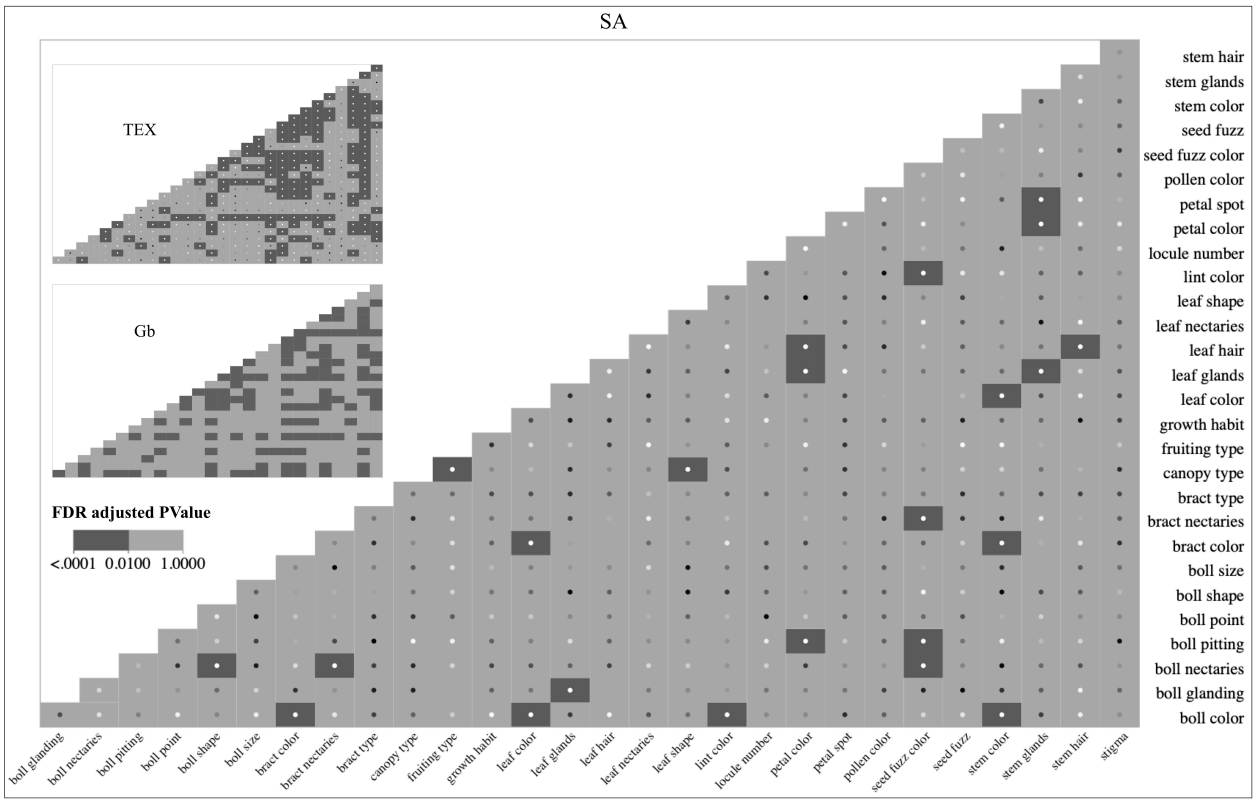


FIGURE 4 | Heat maps of the bivariate descriptor associations were independently evaluated for the SA, TEX, and Gb groups. Larger versions of the TEX and Gb heat figures are in **Supplementary Figures 2, 3**. Interactive heat maps linked to the contingency tables and mosaic plots for each association evaluated are available on-line (https://usda-ars-gbru.github.io/categorical_analysis_cotton/). Dark gray boxes indicate $p < 0.01$. The sample size of accessions for each group is SA: 274, TEX: 471, and Gb: 552.

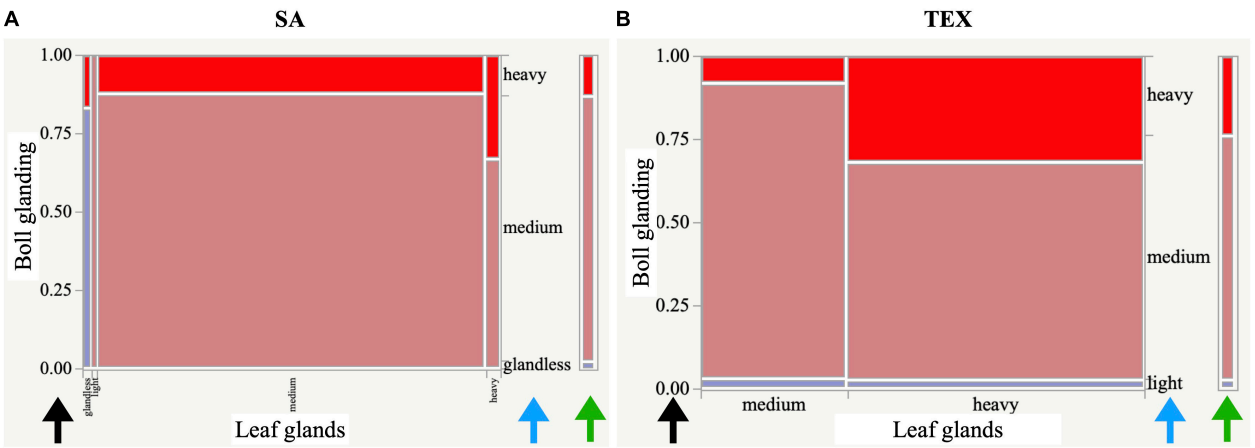


FIGURE 5 | Mosaic plots displaying the degrees of glanding in leaves versus bolls. Plots are shown for **(A)** the SA and **(B)** TEX groups. The plots are divided into rectangles as a stacked bar chart so that the vertical length of each rectangle reflects the proportion of the *Y* variable in each state (blue arrows) of the *X* variable and is a graphical representation of a contingency table. The scale of the vertical axis at left on each plot shows the response probability (black arrows). The whole axis is equivalent to a probability of one, representing the total sample. Fill colors are showing boxes reflecting the phenotype on the *Y*-axis with legend to the right of the figure (green arrows).

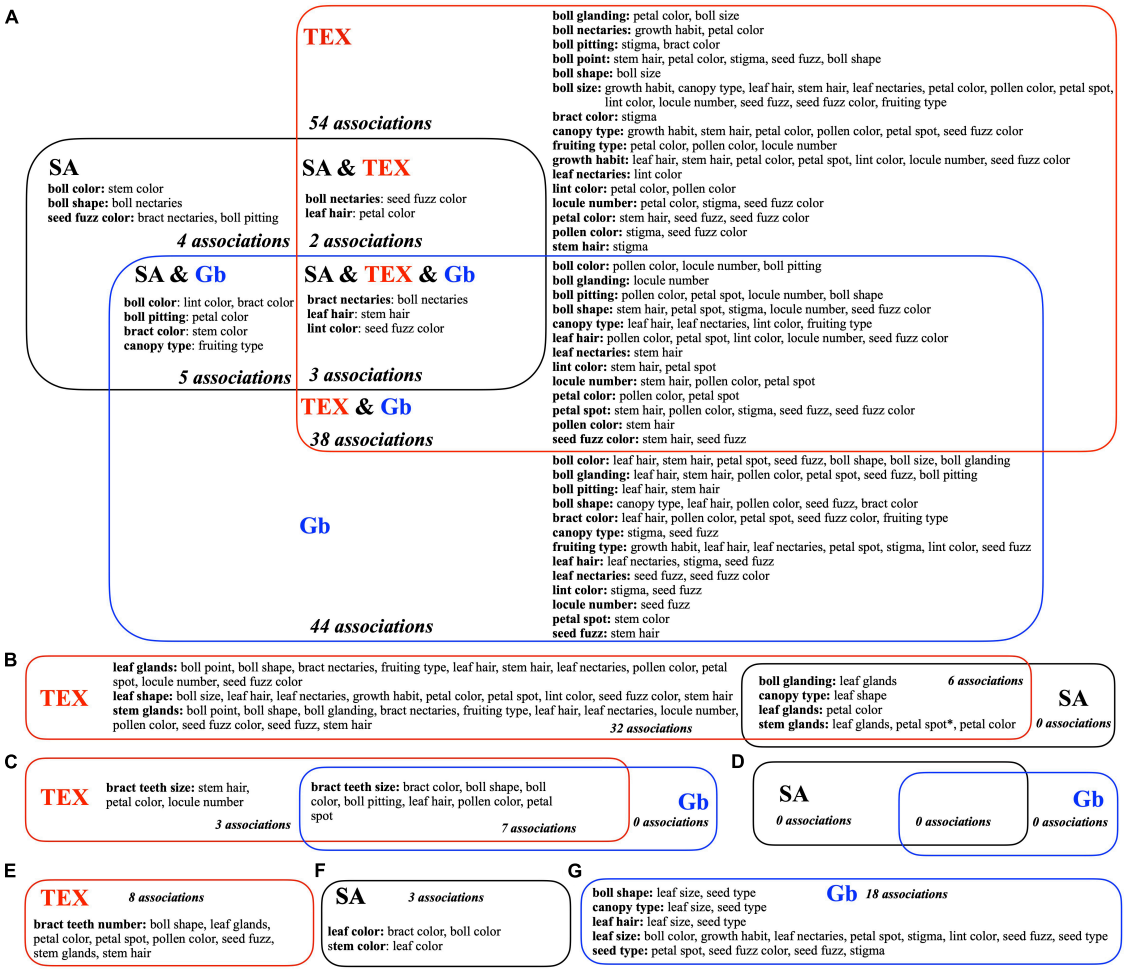


FIGURE 6 | Venn diagrams showing high confidence bivariate descriptor associations within and among three groups of *Gossypium* accessions, SA, TEX, and Gb. Bivariate significance is based on the FDR adjusted *P*-values. Each sector is labeled with the relevant group name(s) and the number of associations it contains. Descriptors within a sector that had high confidence associations ($P < 0.01$) are listed in bold and alphabetical order, with the associated descriptors following in plain text. For example, in the SA and Gb sector **(A)**, the first line indicates two high confidence bivariate associations: boll color with (1) lint color and boll color with (2) bract color. Descriptors listed in plain text may occur in more than a single bold category, but each bivariate descriptor-to-descriptor combination should occur only once in the whole diagram. If an association between two descriptors does not appear, the *p*-value was >0.01 for that comparison. **(A)** Significant categorical descriptors evaluated between the three groups. **(B)** TEX vs. SA. **(C)** TEX vs. Gb. **(D)** SA vs. Gb. **(E)** Only TEX. **(F)** Only SA. **(G)** Only Gb. *In **(B)**, the TEX and SA *stem glands:petal spot* association is the only case where the states of the descriptor states was independently modified for each group (modification shown in **Supplementary Table 6**); results are statistically significant but the descriptor states had different states in the TEX and SA group.

including two descriptor pairs related to gossypol glands, *'boll glanding:leaf glands'* and *'stem glands:leaf glands'* (**Figure 6B**). Between the TEX and the Gb groups, seven diverse plant descriptors were consistently associated with bract teeth size (**Figure 6C**). No high confidence associations were identified in the SA to Gb comparison (**Figure 6D**). Finally, some significant associations occurred only in one group (**Figures 6E–G**).

Unsupervised Clustering Analysis Across Species

Unsupervised clustering analysis was first based on the combined set of SA plus TEX groups (745 total accessions) and the Gb group (552 total accessions) that we had selected for analysis from NCGC in order to determine if the method would generate two species-enriched groups ($k = 2$). In general, the method worked well: the unsupervised (i.e., blind to the NCGC label) k-modes analysis clustered 97.2% of SA plus TEX accessions together and 98.9% of the Gb accessions together (Cluster 2.1 and Cluster 2.2, respectively, in **Table 2**). Only 6 accessions originally labeled as Gb (0.8%) were clustered with the Gh set and only 21 accessions originally labeled as Gh (3.8%) were classified as Gb (**Table 2**). See **Supplementary Table 8** for accession IDs.

Unsupervised Clustering Analysis Across Groups

Unsupervised clustering analysis was then based on a combination of all three groups under analysis to determine if

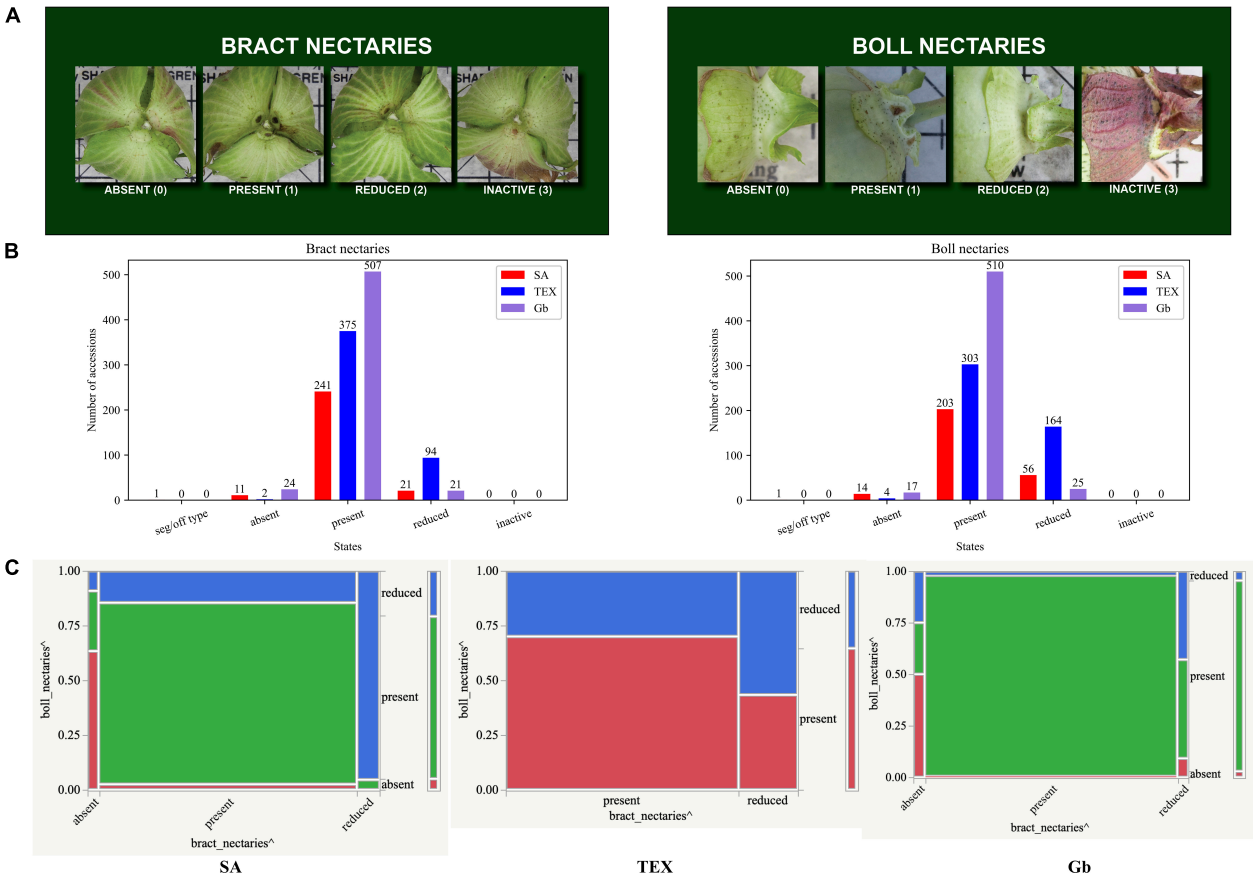


FIGURE 7 | Information related to boll and bract nectaries in SA, TEX, and Gb. The results independently showed a significant association among the three groups (FDR adjusted *P*-values – $P < 0.01$). **(A)** The photos show the rating scale for bract and boll nectaries. **(B)** The two nectary descriptors show a similar distribution of state. (States with ≤5 observations and the seg/off type category were removed prior to bivariate association analysis). **(C)** Mosaic plots from contingency analysis show the relationship among descriptor states for each group evaluated. See **Figure 5** for details on the mosaic plot interpretation.

TABLE 2 | Summary of K-modes unsupervised clustering (k = 2).

	Unsupervised clustering sets	
Original group labels	**Cluster 2.1 (Gh)**	**Cluster 2.2 (Gb)**
SA plus TEX	724	21
Gb	6	546

Two-cluster analysis was designed to group accessions by species. Clusters were arbitrarily numbered based on k-value and unique ID, i.e., 'Cluster 2.1', and the species identifier was assigned afterward based on the majority of pre-labeled accessions in NCGC that it contained.
Silhouette score: 0.33. Accessions analyzed were: SA plus TEX, 745 accessions; and Gb, 552 accessions. The Accessions IDs are reported in ***Supplementary Table 8****.*

TABLE 3 | Summary of K-modes unsupervised clustering (k = 3).

	Unsupervised clustering sets		
Original group labels	**Cluster 3.1 (TEX)**	**Cluster 3.2 (SA)**	**Cluster 3.3 (Gb)**
TEX	308	156	7
SA	21	251	2
Gb	9	0	543

Three-cluster analysis was designed to test for separation between all three groups of accessions. Clusters were arbitrarily numbered based on k-value and unique ID, i.e., 'Cluster 3.1' and the group identifier was assigned afterward based on the majority of pre-labeled accessions in NCGC that it contained.
Silhouette score: 0.22. The Accessions IDs and clusters are reported in ***Supplementary Table 9****.*

the method would separate SA and TEX accessions into two groups while also clustering Gb into a third group (k = 3). The number of accessions analyzed were: 471 for TEX; 274 for SA; and 552 for Gb (see **Supplementary Table 9** for Accessions IDs and clusters.). Results of the clustering are shown in **Table 3**. For the SA group, 91.6% of the accessions were clustered together (Cluster 3.2) and most of the remaining accessions were clustered with the TEX group (Cluster 3.1). A lesser percentage (65.4%) of the TEX group clustered together (Cluster 3.1), with the others (33.1%) grouping with the SA set (Cluster 3.2). Finally, 98.3% of the Gb group clustered together (Cluster 3.3), with a few (1.3%) of the accessions originally labeled as Gb clustering with the TEX group (Cluster 3.1) (**Table 3**).

FIGURE 8 | Multiple correspondence analysis cloud of individuals across groups. The graph shows similarities of the individuals in terms of all categorial descriptor variables. The first two principal components are shown (DIM 1 + DIM 2: 12.4%). Accessions are colored by group for Gb (blue), SA (yellow), and TEX (red); accession index numbers can be found in **Supplementary Tables 2–4**.

Unsupervised Multiple Correspondence Analysis – Clustering Individuals

The MCA "cloud of individuals" appeared to provide similar results to the unsupervised clustering analysis across groups. In the cloud there are 2 notable groups – 1 composed of mostly Gb and 1 composed of SA and TEX accessions. On one hand, most of the Gb accessions are in the negative area between Dim 1 and 2. On the other hand, the SA-TEX cloud shows that most of the SA accessions are in the bottom right area and the TEX are in the top right, though there is a group of TEX accessions located in the SA area (**Figure 8**).

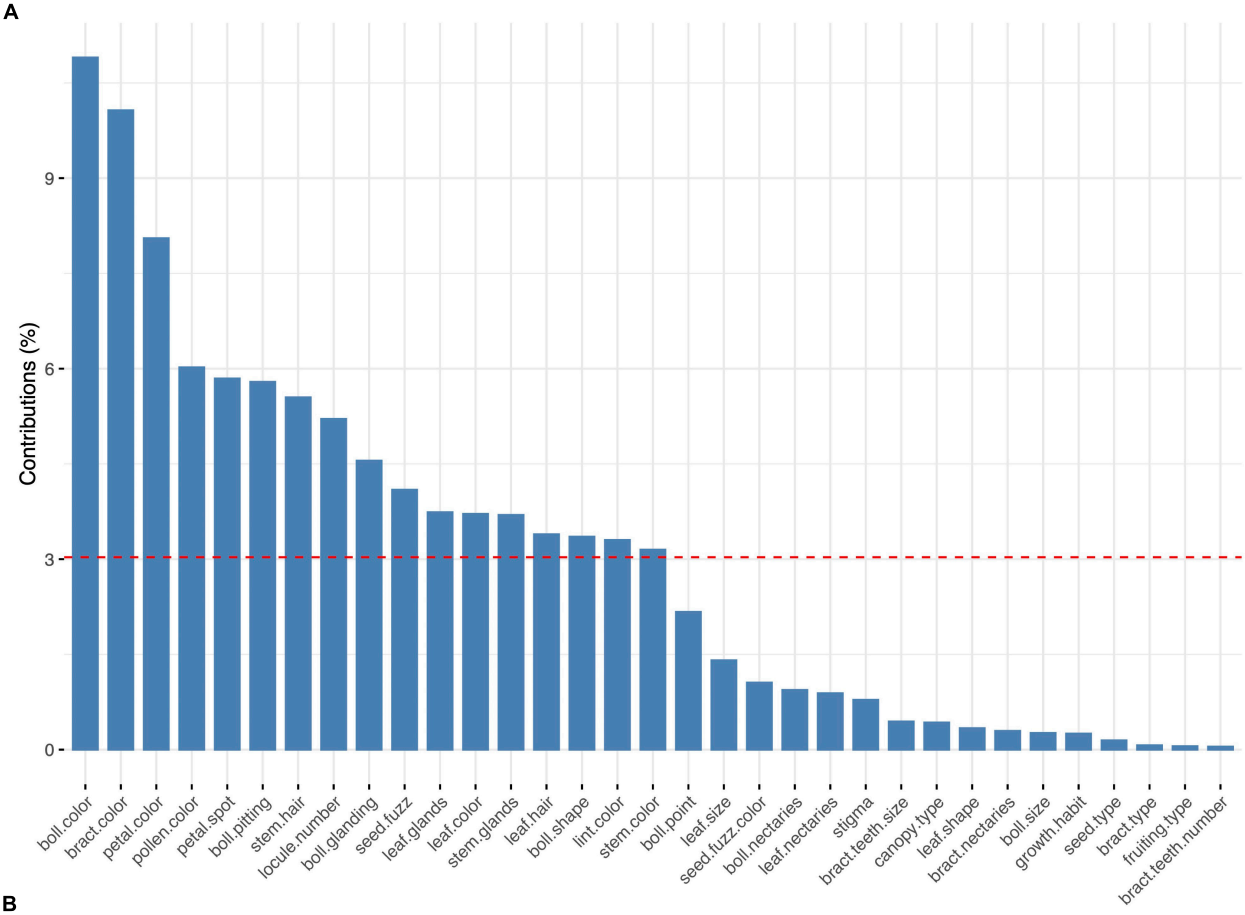
Bivariate Associations Based on Clustering

Three out of the nine sets identified by the unsupervised clustering analysis (k = 3) were reanalyzed using the bivariate association approach. In cluster 3.1, the TEX accessions clustered as TEX (n:308); and for Cluster 3.2, the SA accessions clustered as TEX (n = 156) and the SA accessions clustered as SA (n:251) were processed (IDs in **Supplementary Table 9**), the remaining sets were not evaluated due to the low number of accessions clustered with the exception of Gb to Gb (row by column) in Cluster 3.3, which reports 98% of Gb accessions clustered together and its results are considered highly similar to the results previously shown in **Figures 4**, **6** and reported in https://usda-ars-gbru.github.io/categorical_analysis_cotton/. Generally, the results were not greatly different than the previous bivariate association analysis results, so will not be discussed further, detailed results are available in **Supplementary Figure 7** and **Supplementary Table 10**.

Multiple Correspondence Analysis to Extract Information Content of Descriptors

The calculation of how much each descriptor contributes to the total variation captured by a given Principal component is reported in **Figure 9** for the whole three group set. (Values and corresponding charts are provided for each of the three groups separately, **Supplementary Table 11** and **Supplementary Figure 8**. The cloud of descriptor correlations is reported in **Supplementary Figure 9**). The top contributing descriptors are boll color, bract color, and petal color (**Figure 9A**). The red dashed line indicates the expected average contribution (100% contribution divided by the total number of variables available

A



B

% Variation Explained	# of Descriptors	List of Descriptors
75	13	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands
80	14	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands, leaf hair
85	16	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands, leaf hair, boll shape, lint color
90	17	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands, leaf hair, boll shape, lint color, stem color
95	20	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands, leaf hair, boll shape, lint color, stem color, boll point, leaf size, seed fuzz color
98	24	boll color, bract color, petal color, pollen color, petal spot, boll pitting, stem hair, locule number, boll glanding, seed fuzz, leaf glands, leaf color, stem glands, leaf hair, boll shape, lint color, stem color, boll point, leaf size, seed fuzz color, boll nectaries, leaf nectaries, stigma, bract teeth size

FIGURE 9 | Descriptor contribution across the SA, TEX, and Gb groups. **(A)** Overview of the percent variation contributed to the overall dataset per categorical descriptor. **(B)** Group of categorical descriptors required to capture 75, 80, 85, 90, 95, and 98% of the overall dataset.

in the dataset). Overall, 14, 17, and 24 descriptors can provide 80, 90, and 98 percent of the total variation captured in the overall dataset (**Figure 9B**).

DISCUSSION

Phenotypic descriptors are normally used to catalog plants in the United States GRIN-global germplasm system and for plant registrations. The standardized system that was developed for evaluating phenotypic information for cotton accessions in the NCGC allows for tracking diversity in a unique way while giving stability and evaluation robustness to the germplasm collection data (see Text Footnote 2). The standardized descriptors reflect phenotypic differences between cultivated materials and accessions of other origins that have been deposited in the collection.

Categorical Descriptors Adequately Capture Diversity Between Pima and Upland Cotton

In total, 22% of the total unique accessions in the NCGC were evaluated in this study. The set of accessions evaluated were selected under the criteria that all of the standardized categorical descriptors had been collected for each one because missing information would have reduced statistical power and increased the chances of biased estimates and invalid conclusions. Even though only part of the NCGC cotton collection was analyzed, the high-quality data in the filtered dataset allowed us to draw overall conclusions about phenotypic variation within the SA, TEX, and Gb groups.

The unsupervised two cluster analysis for 33 categorical descriptors adequately separated more than 97% of the Pima and Upland accessions as originally described in the NCGC (k = 2) (**Table 2**). The remaining accessions that were clustered in the opposite group had a combination of descriptors not typical of their previously assigned species in NCGC, which could be due to handling or labeling errors in such a large germplasm collection or to unusual combinations of phenotypes, potentially arising through interspecific crosses or introgression. In certain environments or plant developmental stages, observations of potentially variable traits in hybrids could result in an error or ambiguity in species classification. Overall, this method of clustering accessions based on standardized descriptors can point to accessions within large germplasm banks that need more detailed analysis in order to identify unique and potentially useful genetic combinations and/or to improve the accuracy of the collection records.

The unsupervised three cluster analysis also clearly separated the Gb group, while showing more nuanced outcomes for the groups dominated by Gh accessions: 7.7% of the SA accessions were assigned to the TEX cluster and 33% of the TEX accessions were assigned to the SA cluster. The SA group is referred to as a germplasm breeding reference and contains many cultivars, whereas the TEX group is described as landraces or other tropical materials (Percy et al., 2014). These results are consistent with more extensive breeding to generate Gh cultivars. Early Gh domestication started with ancestors of the landraces that are commonly represented in the TEX group. In addition, more advanced germplasm from Mexico and Central America became an important resource in United States cotton selection and breeding programs beginning in the early 1800s (Moore, 1956; Wendel et al., 2010). Many of these introductions into United States cotton breeding were likely phenotypically quite close to modern Gh cultivars, except for environmentally responsive traits like photoperiodicity that would have been selected against in northerly regions and that were not included in our analysis. Logically, new combinations of phenotypes developed as cotton selection and breeding proceeded over time. Bivariate association analysis may have revealed differences in the composite plant traits between more primitive and advanced accessions as viewed from the cotton breeding perspective (**Table 3**). Genetic information can potentially augment the use of categorial descriptors as described here in further classifying the TEX accessions.

Breeding Has Significantly Impacted the Way Phenotypes Are Associated

Statistical analysis of categorical descriptors collected by the NCGC shows that the breeding process in producing cotton cultivars (SA) has been modifying and reducing the number of significant *'descriptor_1:descriptor_2'* associations compared to the Upland cotton landrace accessions (TEX) and the Pima accessions (Gb). In contrast, most qualitative descriptors have some statistical association with others in the two predominantly Gh groups (SA and TEX) and the Gb group evaluated here (**Figure 6**). Interestingly, the low number of associations in SA is consistent with more extensive breakage of linkages between phenotypes that were originally present in *Gossypium* as compared to TEX which has seen less human manipulation. This is likely due to cotton breeders focusing on many different individual plant traits over time in response to biotic or abiotic stresses. In addition to focusing on specific traits, public breeders have introduced crosses focused on broadening the genetic base of Upland cotton. For example, they have begun evaluating accessions across different environments and looking to exotic or unusual germplasm present in the NCGC for new sources of diversity.

Across the three-group comparison, there were only three shared associations: '*bract nectaries:boll nectaries*', '*leaf hair:stem hair*', and '*lint color:seed fuzz color*' (**Figure 6A**). Therefore, in these particular trait combinations, the association of particular phenotypes across the paired descriptors have not been broken within the accessions analyzed, including Gh and Gb accessions arising through modern breeding. In the case of the '*bract nectaries:boll nectaries*' association, the comparison between groups summarized in the results section suggests that it is uncommon for there to be a difference in presence or absence between boll and bract nectaries in the same accession. This may point to commonalities in the genetic control of nectary formation in both tissues. Despite these persistent pairings, the traits showed a wide variation of states within the range. Such observations are usually explained by polygenic effects

(Waghmare et al., 2005; Hou et al., 2013; Hu et al., 2020). Also, we observed significant bivariate associations between descriptors with no obvious relationships, which could be due to pleiotropic effects when a gene product interacts with multiple others. The currently reported results lead to many future pathways of research to explore the genetic basis of the reported associations, such as the '*canopy type:fruiting type*' significant association case, which is only reported in SA and Gb. Moreover, the canopy type trait reports multiple bivariate associations with other traits in TEX and GB, and independently for TEX, and GB.

The information that we investigated about the diversity of fiber and fuzz color, leaf and stem hairs, nectaries, and boll glands provides additional evidence that the germplasm material serves as a valuable resource in breeding materials for particular traits of interest which are associated with disease resistance, quality, growth habits, and ornamental interests, among others. For example, the data reported about the strong statistical association of presence/absence of nectaries, glands, seed fuzz, and plant hairs allows an interested breeder to identify the accessions with particular physiological conditions showing atypical distribution frequencies to independently explore the biological mechanisms involved in the anomalies of its physiological conditions, such as in the case of '*bract nectaries:boll nectaries*' there are 7 SA lines having present nectaries on bracts but absent on bolls (SA-1009, SA-1034, SA-2242, SA-2861, SA-2870, SA-2925, and SA-3611) and 3 SA lines with the opposite (SA-2946, SA-3570, and SA-3585). These particular trait associations could be targeted specifically for breakdown among elite materials as it potentially indicates there may either be very homogeneous genetic loci shared or in linkage disequilibrium among all the materials which limits potential diversity among other traits of interest shared in those genetic regions or the traits are controlled by some or all the same causal variants. Both factors play a role in this categorical study but exploring those conditions including genetic data could expand the understanding of the mechanisms associated with the traits that breeders could exploit to determine genotype-phenotype patterns. Currently the genomic data is not available for the NCGC but represents a potential future avenue of this research.

We were interested if we could better understand the historical contribution of the materials in this study to cotton breeding and research, which may have targeted use of these materials for certain desired traits as outlined above. The NCGC has tracked the number of total seed requests for each line since 2007 (**Supplementary Tables 2–4**), which should correlate with the utilization of a line in practice. The SA collection has seed request numbers from 0 to 47, averaging 4.5 $\pm$ 5.5 requests per line. The TEX collection has seed request numbers from 0 to 20, averaging 5.8 $\pm$ 3.2 requests per line. The GB collection has seed request numbers from 0 to 57, averaging 4.3 $\pm$ 5.9 requests per line. There were a few major standouts in the SA and GB collections as the most requested lines. In SA, the most requested line is Coker 310 (47 requests); which is an important line from which Coker 312 was selected from, as the most regenerable line of cotton (Trolinder and Goodin, 1987; Bowman et al., 2006). The next most requested lines both have the green lint phenotype, Arkansas Green Lint (42 requests) and Intense Red Green Lint (36 requests), which reflects the interest in cotton that does not require dyeing (Vreeland, 1999). In Gb, the most requested lines are Pima S-6 (57 requests) and Pima S-7 (55 requests), they both have long fiber, good yield and are earlier maturing than most *G. barbadense* lines (Feaster and Turcotte, 1984; Turcotte et al., 1992). The lines have also been studied for their reaction to important diseases such as verticillium wilt and fusarium wilt (Bolek et al., 2005; Wang and Roberts, 2007; Zhu et al., 2021). The third most requested line is Bleak Hall Sea Island (37 requests), an important genetic contributor to the USDA-ARS Pee Dee Breeding Program focused on fiber quality (Campbell et al., 2011). In a field trial of 48 Pima lines, it had the longest fiber length at 37.8 mm (Holladay et al., 2021). The presence of a registration in the Plant Variety Protection (PVP) system often indicates the importance of a line. Of lines studied here, there are only 2 lines that are ex-PVP materials (lines for which a PVP was filed and ex indicating they have passed the time of legal protection), both in the SA collection, Stoneville 907 (*PVP - Stoneville 907, n.d.*) and DP 5409 (*PVP - DP5409, n.d.*). Therefore, it is likely the seed request data provides more data on the importance of the study materials to historical cotton breeding and research.

Resistance-Associated Phenotypes Show Different Patterns of Relationship Among Stoneville Accessions, Texas Accessions, and *Gossypium barbadense*

Gossypol glands play an important role in insect resistance because gossypol is often toxic. The glands are considered direct resistance traits because the plant invests directly in its own defense (Rudgers et al., 2004). In cultivated cotton, the presence and density of glands, which may be found on leaves, stems, and/or bolls, are negatively correlated with the abundance, performance, and/or damage caused by several herbivores (Matthews, 1989; Summy and King, 1992). Results (**Supplementary Figure 6**) showed that both SA and TEX have significant bivariate associations for *'leaf glands:boll glands'* and *'leaf glands:stem glands'*, while *'boll glands:stem glands'* are only significantly associated in the TEX group (**Figure 6**). Most of the accessions in all three groups had at least medium glanding on all three organs (**Figure 3**), which is consistent with a positive impact of glands on defense against insects. Most SA accessions had medium glanding on leaves, stems, and bolls, and rarer cases had glandless leaves and bolls. In contrast, the TEX group contained accessions with glandless bolls accompanied by medium and heavy leaf glands. The majority of TEX accessions had medium glanding in bolls and heavy glanding in leaves and stems. In the Gb group, 98% of the accessions were rated as 'heavy' for the glanding on all three organs (**Figure 3**). Therefore, more extensive breeding in the SA group has led to lesser glanding overall as compared to TEX or Gb. These findings are reasonable from the perspectives of adaptation and evolution because glands provide the plant with natural protection from insects. Thus, losing the glanding trait would be detrimental to overall plant fitness and make it difficult for a breeder to impact glanding. This study is consistent with previous efforts showing the difficulty of

breeding for reduced glanding, potentially indicating alternative breeding methods should be applied where gland modification is the goal (Janga et al., 2019; Gao et al., 2020).

Extrafloral structures such as nectaries reflect indirect resistance mechanisms because the plants invest in interactions with other species (Rudgers et al., 2004). The *'bract nectaries:boll nectaries'* association was significant in all three groups analyzed. The different biological backgrounds of each class and the states observed for descriptors showed differences and similarities in its range trait relationship. The present and reduced states are the most common conditions across the three groups, with absence of bract and boll nectaries only rarely observed among the accessions analyzed. The presence of nectaries could be considered an advantage or disadvantage depending on the natural conditions of the individual in the wild or its use for breeding purposes.

These descriptor traits may be more valuable in ranges where cotton production and specific environmental factors ranges overlap, such as native insect ranges. Assessment of accession geographic collection information, or georeferencing, has led to valuable insights particularly in botanical studies (Swenson et al., 2012; Choudhary et al., 2022). Crop species have a particular difficulty in utilizing geographic data as many accessions were obtained outside of collection expeditions thus contain uninformative or inaccurate geographic data, extensive data filtering would be required to even potentially utilize that data (Feeley and Silman, 2010), but may be worth investigating in the future to potentially add value to the germplasm collection (Volk and Richards, 2011).

Leveraging Germplasm Collection Systems

This analysis expands the use of categorical descriptors normally used for cataloging cotton accessions or germplasm registration. We show that computational and statistical analysis can allow categorical data to be used for illustration and exploration of diversity, trait associations, and similarity in the cotton germplasm collection. The robustness of the analysis is based on the standardized systems developed by the germplasm curators to track multiple phenotypic traits of cotton accessions planted annually in different environments. This research is only based on categorical data and helps to understand the heterogeneity of the cotton accessions present in the collection. This information can be used by the breeding community to integrate new material with desirable traits or unique trait combinations into their breeding programs. While this analysis focused on categorical data as this is the prevalent information available on large numbers of individuals in germplasm collections, a similar strategy can be applied to quantitative data and many of the tools used here are suitable for quantitative data (Lê et al., 2008; Husson et al., 2010; Akay and Yüksel, 2017). As larger quantitative data sets are available for germplasm collections, it will also be possible to combine qualitative and quantitative data for analysis.

We analyzed accessions representing 2 of the over 50 *Gossypium* species represented in the NCGC. The NCGC is only one of 44 collections with over 500,000 unique accessions representing over 10,000 species in the GRIN-global germplasm system [(dataset) USDA Agricultural Research Service, 2015], with different curators all collecting similar categorical descriptor data on the different crop-specific germplasm collections. While determining the specific number of categorical traits necessary to be informative for a collection will be collection specific, the analytical methods and refined insights about the collection demonstrated in this study could be extended to other crops or organisms present in the GRIN-global system. We would suggest a researcher to systematically collect the largest possible set of descriptor traits on a smaller diversity panel or core set of accessions, then use multiple correspondence analysis as outlined here to understand the most informative set of descriptors to collect on the larger collection. A better understanding of germplasm collections will allow for more effective use of these resources and help to safeguard the genetic diversity of agriculturally important plants, which is essential for protecting agriculture in the future (FAO, 2010; Byrne et al., 2018).

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: GitHub (https://usda-ars-gbru.github.io/categorical_analysis_cotton/), GRIN (https://npgsweb.ars-grin.gov/gringlobal/crop?id=547), and Cottongen (https://www.cottongen.org/).

AUTHOR CONTRIBUTIONS

DR-M, AMH-K, JAS, DCJ, and JF conceived the project. JAS, LLH, JL, RGP, and JF managed field locations and data collection, and managed the germplasm collection. DR-M analyzed the data. DR-M, AMH-K, JAS, and CHH synthesized and interpreted the results and wrote the manuscript. All authors edited and approved the manuscript.

FUNDING

This research was funded in part by the United States Department of Agriculture-Agricultural Research Service (USDA-ARS) including ARS project numbers 6066-21310-005-00D, 6066-21000-052-000-D, and 3091-21000-041-000-D. Additional funding supporting DR-M was provided by Cotton Incorporated project 18-274 to AMH-K

ACKNOWLEDGMENTS

The authors wish to thank Wes Malloy and his team at the Cotton Winter Nursery, Tecomán, Mexico, and Alfonso Palafox and his crew at the Cotton Winter Nursery, Liberia, Costa Rica. This research used resources provided by the SCINet project of the

USDA-Agricultural Research Service, ARS project number 0500-00093-001-00-D. The contributions of CHH were aided in part by the USDA National Institute of Food and Agriculture Hatch project 1016883.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.837038/full#supplementary-material

Supplementary Figure 1 | Categorical state distributions for the 33 traits evaluated.

Supplementary Figure 2 | Heatmap of the bivariate descriptor associations for TEX.

Supplementary Figure 3 | Heatmap of the bivariate descriptor associations for Gb.

Supplementary Figure 4 | Leaf hair – stem hair analysis.

Supplementary Figure 5 | Lint color – seed fuzz color analysis.

Supplementary Figure 6 | Boll glanding – leaf glanding – stem glands analysis.

Supplementary Figure 7 | Heatmaps of SA-251, TEX-308, and TEX-156 sets.

Supplementary Figure 8 | Multiple correspondence analysis for SA, TEX, and Gb.

Supplementary Figure 9 | Cloud of descriptor correlations.

Supplementary Table 1 | Standardized descriptors and rating reported by USDA-ARS College Station.

Supplementary Table 2 | SA_mod.xlsx – input dataset for bivariate analysis.

Supplementary Table 3 | TEX_mod.xlsx – input dataset for bivariate analysis.

Supplementary Table 4 | gb_mod_PAG.xlsx – input dataset for bivariate analysis.

Supplementary Table 5 | Categorical descriptors encoded for bivariate and clustering analysis.

Supplementary Table 6 | Modified states in categorical analysis for statistical purposes.

Supplementary Table 7 | Integrated dataset of SA, TEX, and Gb accessions for unsupervised clustering.

Supplementary Table 8 | Two cluster results and IDs identified.

Supplementary Table 9 | Three cluster results and IDs identified.

Supplementary Table 10 | All FDR *p*-values for the SA, TEX, Gb, SA-251, TEX-156, and TEX-308 sets.

Supplementary Table 11 | Values and corresponding charts are provided for SA, TEX, and Gb.

Supplementary Data 1 | Updating and expanding definitions and rating system.

Supplementary Data 2 | "One_hot-ordinal.py" script to transform data to apply the clustering analysis.

Supplementary Data 3 | "MCA_ind_corr_contr_sa_tex_gb.R" script to calculate MCA and plot results.

REFERENCES

Abdi, H., and Valentin, D. (2007). "Multiple correspondence analysis," in *Encyclopedia of Measurement and Statistics*, ed. N. J. Salkind (Thousand Oaks, CA: Sage Publications, Inc), 1–13. doi: 10.4135/9781412952644

Ahmad, S., and Hasanuzzaman, M. (2020). *Cotton Production and Uses: Agronomy, Crop Protection, and Postharvest Technologies*. Berlin: Springer, 1–641. doi: 10.1007/978-981-15-1472-2

Akay, Ö., and Yüksel, G. (2017). Clustering the mixed panel dataset using Gower's distance and k-prototypes algorithms. *Commun. Stat. Simul. Comput.* 47, 3031–3041. doi: 10.1080/03610918.2017.1367806

Allender, C. (2011). The second report on the state of the world's plant genetic resources for food and agriculture. Rome: Food and Agriculture Organization of the United Nations (2010), pp. 370, ISBN 978-92-5-106534-1. *Exp. Agric.* 47, 574–574. doi: 10.1017/S0014479711000275

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* 57, 289–300. doi: 10.1111/J.2517-6161.1995.TB02031.X

Bolek, Y., El-Zik, K. M., Pepper, A. E., Bell, A. A., Magill, C. W., Thaxton, P. M., et al. (2005). Mapping of *Verticillium wilt* resistance genes in cotton. *Plant Sci.* 168, 1581–1590. doi: 10.1016/J.PLANTSCI.2005.02.008

Börner, A., and Khlestkina, E. K. (2019). Ex-situ genebanks—seed treasure chambers for the future. *Russ. J. Genet.* 55, 1299–1305. doi: 10.1134/S1022795419110036

Bowman, D. T., Gutierrez, O. A., Percy, R. G., Calhoun, D. S., and May, O. L. (2006). *Pedigrees of Upland and Pima Cotton Cultivars Released between 1970 and 2005*. Mississippi: Mississippi Agriculture and Forestry.

Byrne, P. F., Volk, G. M., Gardner, C., Gore, M. A., Simon, P. W., and Smith, S. (2018). Sustaining the future of plant breeding: the critical role of the USDA-ARS National Plant Germplasm System. *Crop Sci.* 58, 451–468. doi: 10.2135/CROPSCI2017.05.0303

Cai, Y., Xie, Y., and Liu, J. (2010). Glandless seed and glanded plant research in cotton. A review. *Agron. Sustain. Dev.* 30, 181–190. doi: 10.1051/AGRO/2008024

Camilli, G., and Hopkins, K. D. (1978). Applicability of chi-square to 2 × 2 contingency tables with small expected cell frequencies. *Psychol. Bull.* 85, 163–167. doi: 10.1037/0033-2909.85.1.163

Campbell, B. T., Chee, P. W., Lubbers, E., Bowman, D. T., Meredith, J. R., Johnson, J., et al. (2011). Genetic improvement of the Pee Dee cotton germplasm collection following seventy years of plant breeding. *Crop Sci.* 51, 955–968. doi: 10.2135/CROPSCI2010.09.0545

Campbell, B. T., Saha, S., Percy, R., Frelichowski, J., Jenkins, J. N., Park, W., et al. (2010). Status of the global cotton Germplasm resources. *Crop Sci.* 50, 2198–2198. doi: 10.2135/CROPSCI2009.09.0551ER

Cerda, P., and Varoquaux, G. (2020). Encoding high-cardinality string categorical variables. *IEEE Trans. Knowl. Data Eng.* 34, 1164–1176. doi: 10.1109/TKDE.2020.2992529

Chiu, Y.-F. (2002). Multiple comparisons and multiple tests. Using the SAS system. Peter H. Westfall, Randall D.Tobias, Dror Rom, Russell D. Wolfinger and Yosef Hochberg, SAS Institute, Cary, U.S.A. 2000. No. of pages: xiv + 397. Price: DKK 412.00. ISBN 1-58025-397-0. *Stat. Med.* 21, 1499–1500. doi: 10.1002/SIM.1168

Choudhary, S. B., Gurjar, S. C., Singh, B. K., Singh, D. K., Sharma, H. K., Horo, S., et al. (2022). Morphology and genic-SSRs-based diversity analysis and georeferencing of economic traits in natural populations of Jack (*Artocarpus heterophyllus* Lam.) from Eastern India. *Sci. Hortic.* 295:110852. doi: 10.1016/J.SCIENTA.2021.110852

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ 2 Tests. *Biometrics* 10, 417–451. doi: 10.2307/3001616

de Vos, N. J. (2021). *KModes Categorical Clustering Library*. Available Online at: https://github.com/nicodv/kmodes [accessed October 8, 2021].

di Franco, G. (2016). Multiple correspondence analysis: one only or several techniques? *Qual. Quan.* 50, 1299–1315. doi: 10.1007/S11135-015-0206-0/TABLES/3

FAO (2010). *The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture*. Rome: FAO.

FAO (2017). *The Future of Food and Agriculture: Trends and Challenges*. Rome: FAO.

Feaster, C. V., and Turcotte, E. L. (1984). Registration of pima S-6 cotton. *Crop Sci.* 24, 382–382. doi: 10.2135/CROPSCI1984.0011183X002400020045X

Feeley, K. J., and Silman, M. R. (2010). Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *J. Biogeogr.* 37, 733–740. doi: 10.1111/J.1365-2699.2009.02240.X

Frelichowski, J., and Percy, R. (2015). Germplasm resources collection and management. *Cotton* 57, 45–76. doi: 10.2134/AGRONMONOGR57.2013.0041

Gao, W., Xu, F.-C., Long, L., Li, Y., Zhang, J.-L., Chong, L., et al. (2020). The gland localized CGP1 controls gland pigmentation and gossypol accumulation in cotton. *Plant Biotechnol. J.* 18, 1573–1584. doi: 10.1111/PBI.13323

Gillespie, S., and van den Bold, M. (2017). Agriculture, food systems, and nutrition: meeting the challenge. *Glob. Chall.* 1:1600002. doi: 10.1002/GCH2.201600002

Grover, C. E., Zhu, X., Grupp, K. K., Jareczek, J. J., Gallagher, J. P., Szadkowski, E., et al. (2014). Molecular confirmation of species status for the allopolyploid cotton species, *Gossypium ekmanianum* Wittmack. *Genet. Resour. Crop Evol.* 62, 103–114. doi: 10.1007/S10722-014-0138-X

Holladay, S. K., Bridges, W. C., Jones, M. A., and Campbell, B. T. (2021). Yield performance and fiber quality of Pima cotton grown in the southeast United States. *Crop Sci.* 61, 2423–2434. doi: 10.1002/CSC2.20505

Hou, M., Cai, C., Zhang, S., Guo, W., Zhang, T., and Zhou, B. (2013). Construction of microsatellite-based linkage map and mapping of nectarilessness and hairiness genes in *Gossypium tomentosum*. *J. Genet.* 92, 445–459. doi: 10.1007/S12041-013-0286-3

Hu, W., Qin, W., Jin, Y., Wang, P., Yan, Q., Li, F., et al. (2020). Genetic and evolution analysis of extrafloral nectary in cotton. *Plant Biotechnol. J.* 18, 2081–2095. doi: 10.1111/PBI.13366

Husson, F., Lê, S., and PageÌs, J. (2010). *Exploratory Multivariate Analysis by Example Using R*, Vol. 10. Boca Raton, FL: CRC Press.

Janga, M. R., Pandeya, D., Campbell, L. M., Konganti, K., Villafuerte, S. T., Puckhaber, L., et al. (2019). Genes regulating gland development in the cotton plant. *Plant Biotechnol. J.* 17, 1142–1153. doi: 10.1111/PBI.13044

Kassambara, A. (2016). *Practical Guide to Principal Component Methods in R*. Scotts Valley, CA: CreateSpace.

Kassambra, A. (2022). *Factoextra R Package: Easy Multivariate Data Analyses and Elegant Visualization. Version 1.0.7*. Available online at: https://rpkgs.datanovia.com/factoextra/index.html (accessed March 01, 2022).

Knight, R. L. (1952). The genetics of jassid resistance in cotton. *J. Genet.* 51, 47–66. doi: 10.1007/BF02986704

Kumar, M., Tomar, M., Punia, S., Grasso, S., Arrutia, F., Choudhary, J., et al. (2021). Cottonseed: a sustainable contributor to global protein requirements. *Trends Food Sci. Technol.* 111, 100–113. doi: 10.1016/J.TIFS.2021.02.058

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/JSS.V025.I01

Long, L., Liu, J., Gao, Y., Xu, F. C., Zhao, J. R., Li, B., et al. (2019). Flavonoid accumulation in spontaneous cotton mutant results in red coloration and enhanced disease resistance. *Plant Physiol. Biochem.* 143, 40–49. doi: 10.1016/J.PLAPHY.2019.08.021

Matthews, G. A. (1989). *Cotton Insect Pests and Their Management*. London: Longman Scientific and Technical.

Moore, H. J. (1956). Cotton breeding in the old south. *Agric. Hist.* 30, 95–104.

Nguyen, G. N., and Norton, S. L. (2020). Genebank phenomics: a strategic approach to enhance value and utilization of crop Germplasm. *Plants* 9:817. doi: 10.3390/PLANTS9070817

Nguyen, L. H., and Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* 15:e1006907. doi: 10.1371/JOURNAL.PCBI.1006907

Park, S.-H., Scheffler, J. A., Ray, J. D., and Scheffler, B. E. (2021). Identification of simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) that are associated with the nectariless trait of *Gossypium hirsutum* L. *Euphytica* 217:78. doi: 10.1007/S10681-021-02799-8

Park, S.-H., Scheffler, J., Scheffler, B., Cantrell, C. L., and Pauli, C. S. (2019). Chemical defense responses of upland cotton, *Gossypium hirsutum* L. to physical wounding. *Plant Direct* 3:e00141. doi: 10.1002/PLD3.141

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.1080/13696998.2019.1666854

Percy, R. G., Frelichowski, J. E., Arnold, M. D., Campbell, T. B., Dever, J. K., Fang, D. D., et al. (2014). "The U.S. national cotton Germplasm collection – its contents, preservation, characterization, and evaluation," in *World Cotton Germplasm Resources*, ed. I. Y. Abdurakhmonov (Rijeka: InTech), 167–201. doi: 10.5772/58386

Percy, R. G. G., and Kohel, R. J. J. (1999). "Qualitative genetics," in *Cotton: Origin, History, Technology, and Production*, eds W. C. Smith and J. T. Cothren (Hoboken, NJ: John Wiley & Sons), 319–360.

Postman, J., Hummer, K., Bretting, P., Kinard, G., Bohning, M., Emberland, G., et al. (2010). GRIN-global: an international project to develop a global plant Genebank information management system. *Acta Hortic.* 859, 49–56. doi: 10.17660/ACTAHORTIC.2010.859.4

PVP - DP5409 (n.d.). Available Online at: https://apps.ams.usda.gov/CMS/AdobeImages/009300189.pdf [accessed March 13, 2022].

PVP - Stoneville 907 (n.d.). Available Online at: https://apps.ams.usda.gov/CMS/AdobeImages/009200016.pdf [accessed March 13, 2022].

Ramankutty, N., Mehrabi, Z., Waha, K., Jarvis, L., Kremen, C., Herrero, M., et al. (2018). Trends in global agricultural land use: implications for environmental health and food security. *Annu. Rev. Plant Biol.* 69, 789–815. doi: 10.1146/ANNUREV-ARPLANT-042817-040256

Rudgers, J. A., Strauss, S. Y., and Wendel, J. F. (2004). Trade-offs among anti-herbivore resistance traits: insights from *Gossypieae* (Malvaceae). *Am. J. Bot.* 91, 871–880. doi: 10.3732/AJB.91.6.871

Summy, K. R., and King, E. G. (1992). Cultural control of cotton insect pests in the United States. *Crop Prot.* 11, 307–319. doi: 10.1016/0261-2194(92)90055-A

Swenson, N. G., Enquist, B. J., Pither, J., Kerkhoff, A. J., Boyle, B., Weiser, M. D., et al. (2012). The biogeography and filtering of woody plant functional diversity in North and South America. *Glob. Ecol. Biogeogr.* 21, 798–808. doi: 10.1111/J.1466-8238.2011.00727.X

Tian, Z., Wang, J.-W., Li, J., and Han, B. (2021). Designing future crops: challenges and strategies for sustainable agriculture. *Plant J.* 105, 1165–1178. doi: 10.1111/TPJ.15107

Trolinder, N. L., and Goodin, J. R. (1987). Somatic embryogenesis and plant regeneration in cotton (*Gossypium hirsutum* L.). *Plant Cell Rep.* 6, 231–234. doi: 10.1007/BF00268487

Turcotte, E. L., Percy, R. G., and Feaster, C. V. (1992). Registration of "Pima S-7" American Pima cotton. *Crop Sci.* 32:1291. doi: 10.2135/cropsci1992.0011183x003200050047x

UPOV-Council (2019). *Trial Design and Techniques Used in the Examination of Distinctness, Uniformity, and Stability (Document TG/8). Associated Document to the General Introduction to the Examination of Distinctness, Uniformity and Stability and the Development of Harmonized Descriptions of New Varieties of Plants (Document TG/1/3)*. Available Online at: https://www.upov.int/edocs/tgpdocs/en/tgp_8.pdf [accessed November 1, 2019].

USDA Agricultural Research Service (2015). *Germplasm Resources Information Network (GRIN)* (dataset). Beltsville, MD: USDA/ARS. doi: 10.15482/USDA.ADC/1212393

Volk, G. M., and Richards, C. M. (2011). Integration of georeferencing, habitat, sampling, and genetic data for documentation of wild plant genetic resources. *HortScience* 46, 1446–1449. doi: 10.21273/HORTSCI.46.11.1446

Vreeland, J. M. (1999). The revival of colored cotton on JSTOR. *Sci. Am.* 280, 112–118. doi: 10.1038/scientificamerican0499-112

Waghmare, V. N., Rong, J., Rogers, C. J., Pierce, G. J., Wendel, J. F., and Paterson, A. H. (2005). Genetic mapping of a cross between *Gossypium hirsutum* (cotton) and the Hawaiian endemic, *Gossypium tomentosum*. *Theor. Appl. Genet.* 111, 665–676. doi: 10.1007/S00122-005-2032-6

Wang, C., and Roberts, P. A. (2007). A *Fusarium wilt* resistance gene in *Gossypium barbadense* and its effect on root-knot nematode-wilt disease complex. *Phytopathology* 96, 727–734. doi: 10.1094/PHYTO-96-0727

Wallace, T. P., Bowman, D., Campbell, B. T., Chee, P., Gutierrez, O. A., Kohel, R. J., et al. (2008). Status of the USA cotton Germplasm collection and crop vulnerability. *Genet. Resour. Crop Evol.* 56, 507–532. doi: 10.1007/S10722-008-9382-2

Watson, K. B. (2014). "Categorical data analysis," in *Encyclopedia of Quality of Life and Well-Being Research*, ed. A. C. Michalos (Dordrecht: Springer), 601–604. doi: 10.1007/978-94-007-0753-5_291

Wendel, J. F., Brubaker, C. L., and Seelanan, T. (2010). "The origin and evolution of *Gossypium*," in *Physiology of Cotton*, eds J. M. Stewart, D. M. Oosterhuis, J. J. Heitholt, and J. R. Mauney (Dordrecht: Springer), 1–18. doi: 10.1007/978-90-481-3195-2_1

White, G. A., Shands, H. L., and Lovell, G. R. (2011). History and operation of the national plant Germplasm system. *Plant Breed. Rev.* 7, 5–56. doi: 10.1002/9781118061046.CH1

Wilkes, G., and Williams, J. T. (2008). Current status of crop plant Germplasm. *Crit. Rev. Plant Sci.* 1, 133–181. doi: 10.1080/07352688309382175

Yuan, D., Grover, C. E., Hu, G., Pan, M., Miller, E. R., Conover, J. L., et al. (2021). Parallel and intertwining threads of domestication in allopolyploid cotton. *Adv. Sci.* 8:2003634. doi: 10.1002/ADVS.202003634

Zeng, L., Stetina, S. R., Erpelding, J. E., Bechere, E., Turley, R. B., and Scheffler, J. (2018). History and current research in the USDA-ARS cotton breeding program at Stoneville, MS. *J. Cotton Sci.* 22, 24–35.

Zhang, Z., Wang, P., Luo, X., Yang, C., Tang, Y., Wang, Z., et al. (2019). Cotton plant defence against a fungal pathogen is enhanced by expanding BLADE-ON-PETIOLE1 expression beyond lateral-organ boundaries. *Commun. Biol.* 2:238. doi: 10.1038/s42003-019-0468-5

Zhao, C., Zhang, Y., Du, J., Guo, X., Wen, W., Gu, S., et al. (2019). Crop phenomics: current status and perspectives. *Front. Plant Sci.* 10:714. doi: 10.3389/FPLS.2019.00714

Zhou, M., Zhang, C., Wu, Y., and Tang, Y. (2013). Metabolic engineering of gossypol in cotton. *Appl. Microbiol. Biotechnol.* 97, 6159–6165. doi: 10.1007/S00253-013-5032-5

Zhu, Y., Abdelraheem, A., Cooke, P., Wheeler, T., Dever, J. K., Wedegaertner, T., et al. (2021). Comparative analysis of infection process in Pima cotton differing in resistance to *Fusarium wilt* caused by *Fusarium oxysporum* f. sp. *vasinfectum* race 4. *Phytopathology* HYTO05210203R. doi: 10.1094/PHYTO-05-21-0203-R

Author Disclaimer: 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.

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.

frontiers | Frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 May 2022
doi: 10.3389/fpls.2022.882136

RNA-Seq Provides Novel Genomic Resources for Noug (*Guizotia abyssinica*) and Reveals Microsatellite Frequency and Distribution in Its Transcriptome

Adane Gebeyehu[1,2,3], Cecilia Hammenhag[1], Kassahun Tesfaye[2,3], Ramesh R. Vetukuri[1], Rodomiro Ortiz[1] and Mulatu Geleta[1]*

[1] Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma, Sweden, [2] Ethiopian Biotechnology Institute, Addis Ababa, Ethiopia, [3] Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia

OPEN ACCESS

Edited by:
Andrés J. Cortés, Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

Reviewed by:
Jan Graffelman, Universitat Politecnica de Catalunya, Spain
Weihua Qiao, Chinese Academy of Agricultural Sciences (CAAS), China

***Correspondence:**
Adane Gebeyehu
adane.gebeyehu.demissie@slu.se; adyamrot@gmail.com

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

Received: *23 February 2022*
Accepted: *23 March 2022*
Published: *11 May 2022*

Citation:
Gebeyehu A, Hammenhag C, Tesfaye K, Vetukuri RR, Ortiz R and Geleta M (2022) RNA-Seq Provides Novel Genomic Resources for Noug (Guizotia abyssinica) and Reveals Microsatellite Frequency and Distribution in Its Transcriptome. Front. Plant Sci. 13:882136. doi: 10.3389/fpls.2022.882136

Genomic resources and tools are essential for improving crops and conserving their genetic resources. *Guizotia abyssinica* (noug), an outcrossing edible oilseed crop, has highly limited genomic resources. Hence, RNA-Seq based transcriptome sequencing of 30 noug genotypes was performed to generate novel genomic resources and assess their usefulness. The genotypes include self-compatible and self-incompatible types, which differ in maturity time, photoperiod sensitivity, or oil content and quality. RNA-Seq was performed on Illumina HiSeq 2500 platform, and the transcript was reconstructed *de novo*, resulting in 409,309 unigenes. The unigenes were characterized for simple sequence repeats (SSRs), and served as a reference for single nucleotide polymorphism (SNP) calling. In total, 40,776 SSRs were identified in 35,639 of the 409,309 unigenes. Of these, mono, di, tri, tetra, penta and hexanucleotide repeats accounted for 55.4, 20.8, 21.1, 2.3, 0.2, and 0.2%, respectively. The average G+C content of the unigenes and their SSRs were 40 and 22.1%, respectively. The vast majority of mononucleotide repeat SSRs (97%) were of the A/T type. AG/CT and CCA/TGG were the most frequent di and trinucleotide repeat SSRs. A different number of single nucleotide polymorphism (SNP) loci were discovered in each genotype, of which 1,687 were common to all 30 genotypes and 5,531 to 28 of them. The mean observed heterozygosity of the 5,531 SNPs was 0.22; 19.4% of them had polymorphism information content above 0.30 while 17.2% deviated significantly from Hardy-Weinberg equilibrium ($P < 0.05$). In both cluster and principal coordinate analyses, the genotypes were grouped into four major clusters. In terms of population structure, the genotypes are best represented by three genetic populations, with significant admixture within each. Genetic similarity between self-compatible genotypes was higher, due to the narrow genetic basis, than that between self-incompatible genotypes. The genotypes that shared desirable characteristics, such as early maturity, and high oil content were found to be genetically diverse, and hence

superior cultivars with multiple desirable traits can be developed through crossbreeding. The genomic resources developed in this study are vital for advancing research in noug, such as genetic linkage mapping and genome-wide association studies, which could lead to genomic-led breeding.

Keywords: *de novo* transcriptome assembly, G+C content, genetic variation, self-compatibility, SNPs, SSR, unigenes

INTRODUCTION

Noug (*Guizotia abyssinica*) is an edible oilseed crop indigenous to Ethiopia, where it was originated, domesticated and genetically diversified. It is an annual diploid crop with 2n = 30 chromosomes (Dagne, 1994) exhibiting a strict outcrossing reproductive mechanism with honeybees as major pollinators due to its homomorphic self-incompatibility (Geleta et al., 2002; Geleta, 2007; Geleta and Bryngelsson, 2010). It is among major edible oilseed crops grown in Ethiopia, both in terms of acreage and production volume, where 26% of the produce is consumed locally (Geleta and Ortiz, 2013; Ethiopian Institute of Agricultural Research [EIAR], 2017). It is also cultivated to some extent in other African countries that include Sudan, Malawi and Uganda (Geleta and Ortiz, 2013; Gebeyehu et al., 2021). Apart from Africa, it is cultivated in India as a minor oilseed crop, as well as in Bangladesh, the Caribbean, and the United States, however to a much lesser extent (Geleta and Ortiz, 2013).

Genetic diversity in crops refers to the genetic variation within and between individual plants, landrace populations, and cultivars, which results from mutation, recombination, introgression, natural and artificial selection, and adaptation to diverse environments. A crop's genetic diversity is typically greatest in areas where it was domesticated, originated, or has wild relatives (Geleta and Ortiz, 2013). This diversity plays a key role in the crop's ability to adapt to climate change and withstand new pests, as well as to increase its productivity and quality. Since Ethiopia is its center of origin and diversity, noug cultivated in the country is inherently diverse with high genetic potential for improvement (Geleta et al., 2007, 2008; Petros et al., 2007; Dempewolf et al., 2010; Mengistu et al., 2020; Tsehay et al., 2020). However, the genetic potential of this crop has not been widely exploited, and only a few modestly improved cultivars have been released (Alemaw and Alamayehu, 1997). Among the major constraints are strict self-incompatibility, which requires abundant availability of insect pollinators, an indeterminate growth habit that leads to seed loss due to shattering, lodging, low response to management and inputs, and pests (including various pathogens, insects and parasitic weeds).

The process of cultivar development for a crop begins with selecting genetic material with desirable traits. For efficient selection of genetic material for breeding, understanding the genetic variation within a crop's gene pool is vitally important using DNA markers. Thus, it is imperative that genome-wide markers be developed and utilized in order to identify and manage genetic diversity within a crop's gene pool and to determine genetic factors determining desirable traits. To interpret the functional elements of a genome, it is essential to understand its transcriptome, which include sequence variation in their mRNA transcripts (Wang et al., 2009). As transcriptome markers represent the expressed parts of a genome, they are a better choice than genomic markers for aforementioned applications. To this end, a limited number of transcriptome sequences have been assembled for noug (Dempewolf et al., 2010; Hodgins et al., 2014; Tsehay et al., 2020), and based on these, simple sequence repeat (SSR) markers and single nucleotide polymorphism (SNP) markers have been developed (Dempewolf et al., 2010; Tsehay et al., 2020). However, these genomic resources are insufficient for use in different applications including population genetics analyses for conservation; genome-wide association studies (GWAS) as well as for enabling genomics-led breeding. Hence, the development of additional genomic resources for noug is vitally important.

RNA-Seq (RNA sequencing) is the most advanced method of profiling transcriptomes, which relies on next-generation sequencing methods for high-throughput (Wang et al., 2009). The capability of detecting sequence variations, such as Indels and SNPs in transcribed genomic regions are among the key advantages of RNA-Seq (Cloonan et al., 2008). Additionally, the unigenes obtained after transcriptome assembly can be used in the development of other markers, such as SSRs. The aims of this study were to use RNA-Seq for transcriptome sequencing of diverse genotypes of noug for the development of new genomic resources for their various applications, characterize the SSRs in the unigenes, and assess the usefulness of the novel SNP markers *via* genetic diversity analyses of the genotypes used.

MATERIALS AND METHODS

Plant Material

Thirty phenotypically diverse noug genotypes were used in this study (**Supplementary Table 1**). Most of the genotypes were selected from breeding populations bred for desirable traits such as self-compatibility, early maturity, less-sensitivity to photoperiod, as well as high oil or increased oleic acid contents (Geleta and Bryngelsson, 2010; Geleta et al., 2011; Geleta and Ortiz, 2013). Other genotypes were selected from landrace populations based on their distinct differences in one or more traits from those that were already selected (**Supplementary Table 1**). Twelve of the 30 genotypes are self-compatible although to a different extent, whereas the remaining eighteen are strictly self-incompatible. In terms of maturity time, the source populations varied from very-early to very-late types. For three of the 30 genotypes, the source populations were able to flower when the photoperiod was above 12 h. The average oil content of

the source populations varied from 30 to 45% of dry seed weight. As opposed to the other source populations, four have oleic acid content above 10%, although the level depends primarily on environmental temperature (**Supplementary Table 1**).

Planting, Sampling and RNA Extraction

The 30 genotypes were planted using 1.5 L plastic pots filled with soil in a greenhouse at the Swedish University of Agricultural Sciences (SLU, Alnarp, Sweden) for RNA extraction. Four weeks after planting, leaf tissue was collected separately from individual plants of each genotype in 15 ml falcon tubes and snap-frozen in liquid nitrogen and then stored at −80°C until used for RNA extraction. For each sample, the total RNA was extracted from approximately 100 mg leaf tissue using the RNeasy Plant Mini Kit (#74904, QIAGEN) according to the manufacturer's protocol, followed by DNase treatment with Ambion Turbo DNA-Free Kit (#AM1907, Thermo Fisher Scientific, CA, United States) as described in Kalyandurg et al. (2021). The extracted RNA quality and quantity were assessed using an Agilent Bioanalyzer 2100 system (Agilent, Technologies, CA, United States), NanoDrop ND-1000 spectrophotometer (Saveen Werner, Sweden), and agarose gel electrophoresis. Then, high-quality RNA samples were sent to CD Genomics (New York, United States) for RNA-Seq analysis. Upon arrival, the samples were further monitored on 1% agarose gels for degradation and contamination, purity checked using the NanoPhotometer spectrophotometer (IMPLEN, CA, United States), concentration measured using the Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life Technologies, CA, United States), integrity assessed using the RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, United States).

Library Preparation, Clustering and Sequencing

The NEBNext UltraTM RNA Library Prep Kit for Illumina (NEB, United States) was used to create sequencing libraries from 1.5 μg of RNA per sample, according to the manufacturer's instructions, and index codes were added to assign sequences to each sample. An AMPure XP system (Beckman Coulter, Beverly, United States) was used to purify the library fragments to facilitate preferential selection of cDNA fragments with a length of 150–200 bp. Following adapter ligation to the size-selected fragments and polymerase chain reaction (PCR), the AMPure XP system was used to purify the amplified products, and then library quality was assessed using the Agilent Bioanalyzer 2100 system. This was followed by the clustering of the index-coded samples on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumia) as per the manufacturer's instructions. The clusters were then sequenced on the Illumina HiSeq 2500 platform, and paired-end reads were generated.

Data Quality Control, *de novo* Transcript Assembly and Splicing, and SSR Identification

The Illumina Hiseq data was translated to sequenced reads through base calling, and a FASTQ file containing sequenced reads and quality information was created from the raw data for each sample. A series of methods was applied to filter the raw sequencing reads to obtain high quality data for subsequent analysis. First, the raw reads in FASTQ format were processed using in-house python scripts, and reads containing adapter and ploy-N were removed to obtain clean reads. The Phred quality scores of the clean reads were then calculated, and those with Phered quality scores below 30 (error rate greater than 0.1%) were removed. The remaining high-quality reads were used for downstream analyses.

Since noug does not have a reference genome, *de novo* transcript reconstruction was done using Trinity software package (Grabherr et al., 2011). For this, read1 files containing high-quality reads for each of the 30 samples were merged into a single read1 file, and similarly the read2 files of the 30 samples were merged into a single read2 file. The merged read1 and read2 files were then used for transcript assembly and splicing using Trinity, by setting max_kmer_cov to 2 and all other parameters to default. Following length distribution analysis, the longest spliced transcript for each gene was identified as a unigene and used as a reference sequence in subsequent analyses. This resulted in 409,309 unigenes with a G+C content of 40%, which were used as reference for SNP calling. A web-based microsatellite identification tool MISA-web (Beier et al., 2017[1]) was used to identify simple sequence repeats (SSRs) within the unigenes using the default setting. The minimum number of repeats was set to ten for mononucleotide repeats, to six for dinucleotide repeats and to five for tri, tetra, penta and hexanucleotide repeats.

SNP Calling and Further Processing

As the first step of SNP calling, the BWA v.0.7.4 short read aligner was used to align the high-quality clean reads of each sample to the reference transcripts (Li and Durbin, 2009). Then, SAMtools v0.1.18 (Li et al., 2009) and Picard-tools v1.41 software packages were used for sorting, indexing, removing duplicates, and merging the BAM alignment results of each sample. On the merged BAM files, the Genome Analysis Toolkit (GATK; McKenna et al., 2010) was used for base-quality score calibration, and SNP calling, and genotyping for each sample was performed by using standard filtering parameters or variant quality score calibration according to GATK's Best Practice recommendations (DePristo et al., 2011; Van der Auwera and O'Connor, 2020). The VCF files of the samples were then merged and the shared SNP loci were filtered using BCFtools (Danecek et al., 2021).

Statistical Analysis

Different statistical programs were used to estimate genetic diversity parameters and indices for each genotype across loci and for each locus across genotypes. GenAlEx version 6.5 software (Peakall and Smouse, 2006) was used for the analysis of mean values of observed number of alleles (Na), observed heterozygosity (Ho), number of private alleles (NPA), percent polymorphic loci (PPL) for each genotype, Nei's standard genetic distance (GD) and GD-based principal coordinate analysis (PCoA) to display the genetic relationship between

[1]http://pgrc.ipk-gatersleben.de/misa/misa.html

the noug genotypes based on both SNP data sets. Pairwise GD matrices were also used for neighbor joining (NJ)-based cluster analysis using the MEGA7 program (Kumar et al., 2016). The polymorphism information content of each SNP locus was calculated in accordance with Hildebrand et al. (1992). Arlequin v. 3.5.2.2 (Excoffier and Lischer, 2010) was used to perform the exact test of Hardy-Weinberg equilibrium (using 1,000,000 steps in the Markov chain and 100,000 dememorization steps), and calculate pairwise F_{ST} and mean number of pairwise differences between and within genotypes and groups. To generate heatmaps of these parameters, a console version of the R statistical package (Rcmd) incorporated into the Arlequin software was used. A Bayesian statistics based population genetic structure analysis was conducted using STRUCTURE software version 2.3.4 (Pritchard et al., 2000). The analysis was conducted using an admixture model for different number of clusters (K) using 100,000 burn-in periods and 200,000 Markov chain Monte Carlo (MCMC) chain iterations, with K ranging from two to ten and twenty replications at each K. A further analysis of the results was performed with the STRUCTURESELECTOR (Li and Liu, 2018) program to determine the number of clusters (genetic populations) according to the Puechmaille (2016) method, and to visualize the population structure using CLUMPAK (Kopelman et al., 2015) integrated into STRUCTURESELECTOR.

RESULTS

SSR Identification and Characterization

The analysis of 409,309 unigenes using MISA-web for detecting SSRs resulted in 40,776 SSRs (**Table 1**). These SSRs were detected in 35,639 unigenes (8.7% the total unigenes), of which 4,269 had more than one SSR (1% of the total unigenes, or 12% of the unigenes containing SSRs). Some of these SSRs were separated by less than 100 bases and hence formed compound SSRs. Counting SSRs forming a compound SSR as one, the total number of SSRs was 38,011, of which 2,380 were compound SSRs (**Table 1** and **Supplementary Table 2**). Among the 40,776 separate SSRs identified, mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats accounted for 55.4, 20.8, 21.1, 2.3, 0.2, and 0.2%, respectively (**Table 1** and **Figure 1**). In all cases, the lowest number of repeats accounted for the highest proportion. Among the mononucleotide repeat SSRs, 50.9% had a repeat of ten whereas 37.7% of dinucleotide repeat SSRs had a repeat of six. In the case of tri, tetra, penta and hexanucleotide repeats, a repeat of five accounted for 59.6, 76.9, 79.6, and 50.6%, respectively (**Figure 1**). In general, the longer a given SSR motif gets, the less frequent it becomes. The G+C contents of mono, di, tri, tetra, penta and hexanucleotide SSRs were 3.9, 33.3, 43.1, 25.5, 41.5, and 41.0%, respectively. Whereas, all SSRs together had a G+C content of 22.2% (**Figure 2A**).

The SSRs were further analyzed considering sequence complementarity (**Figure 2**). Among the mononucleotide repeat SSRs, the vast majority (97.2%) were A/T type whereas C/G type accounted for only 2.8% (**Figure 2**). The most and least common dinucleotide repeat SSRs were AG/CT (19.6%) and CG/CG (0.09%), respectively. CCA/TGG, ACC/GGT, and ATC/GAT were the top three most common trinucleotide repeat SSRs, accounting for 10.6, 10.1, and 9.2% of the total trinucleotide repeat SSRs, respectively. Among the tetranucleotide repeat SSRs, AACA/TGTT was by far the most frequent (36.7%), followed by AAAC/GTTT (12.4%). The most common pentanucleotide repeats were AAACC/GGTTT, ATCCA/TGGAT, and CCAAA/TTTGG (12, 11, and 11%, respectively). The frequency of different types of hexanucleotide repeats ranged from one to five, with AGATGA/TCATCT being the most common (**Figure 2**).

The SNP Markers

The number of high-quality SNPs discovered in each sample that met all filtering criteria ranged from 80,653 (in genotype Ga02.02) to 334,828 (in genotype Ga09.03) (Data not shown). Among the SNPs discovered in each genotype, 1,687 of them were shared among the 30 genotypes. In comparison, excluding two of the samples (Ga02.02 and Ga101B.m) that shared the least number of SNP loci with the others resulted in 5,531 SNP loci shared by the 28 remaining samples (**Figure 3** and **Supplementary Table 3**). Both SNP datasets were used for further analyses and the results were compared. Out of the 5,531 SNP loci, 1,500 (27%) were monomorphic across the 28 genotypes whereas 542 of the 1,687 SNP loci (32.1%) were monomorphic across the 30 genotypes (**Figure 3**). Thus, the number of polymorphic SNPs was 4,031 for the 28 genotypes and 1,145 for the 30 genotypes.

Among the SNP loci shared by the 28 and 30 genotypes, 1,074 (19.4%) and 200 (11.9%) loci had a polymorphism information content (PIC) of above 0.30, respectively (**Figure 3** and **Supplementary Table 3**), and hence are highly informative. Under the assumption that the genotypes constitute a random sample of a single population, the HWE test revealed that 953

TABLE 1 | Summary information about the simple sequence repeat (SSR) analysis.

SSR Analysis	No. of genes	Percentage (%)
Total number of sequences examined (TNSE)	409,309	100[a]
Total size of examined sequences (bp)	204,196,448	
Number of SSR containing sequences	35,639	8.7[a]
Number of sequences containing more than one SSRs	4,269	1.0[a]
Total number of identified SSRs (TNIS)	40,776	100[b]
Number of mononucleotide repeat SSRs	22,582	55.4[b]
Number of dinucleotide repeat SSRs	8,487	20.8[b]
Number of trinucleotide repeat SSRs	8,589	21.1[b]
Number of tetranucleotide repeat SSRs	938	2.3[b]
Number of pentanucleotide repeat SSRs	93	0.2[b]
Number of hexanucleotide repeat SSRs	87	0.2[b]
Total number of SSRs (TNS)*	38,011	100[c]
Number of SSRs present in compound formation	2,380	6.3[c]

Number of repeats considered for mononucleotide SSRs: ≥ 10.
Number of repeats considered for dinucleotide SSRs: ≥ 6.
Number of repeats considered for tri, tetra, penta and hexanucleotide repeats: ≥ 5.
Maximal number of bases interrupting two SSRs in a compound SSR: = 100.
**Compound SSRs were counted as single SSRs unlike the case of TNIS.*
[a] %age of TNSE; [b] = %age of TNIS; [c] = %age of TNS.

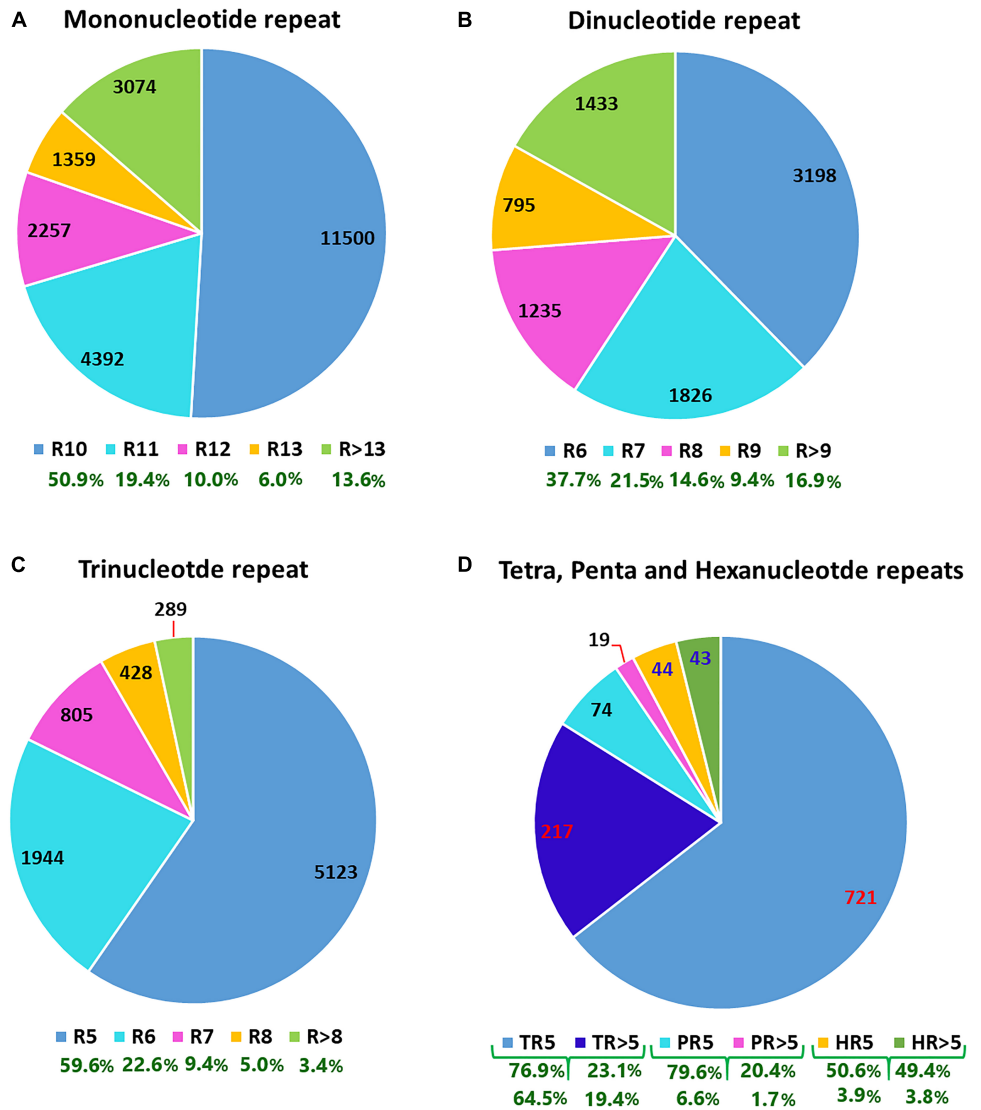


FIGURE 1 | P-charts depicting the number (values in the pie charts) and percentage (values shown below the chart keys in green) of different simple sequence repeat motifs across the noug unigenes that were classified based on the number of times they were repeated. In the chart keys of panels **(A–C)**, the numbers following "R" denote the number of times the corresponding repeat motifs were repeated. In the chart keys of **(D)**, T, P and H refer to Tetra, Penta and Hexanucleotide repeats. ">" indicates that the SSR motif was repeated more times than the specified number.

loci (17.2% of the 5,531 loci) and 167 loci (9.9% of the 1,687 loci) showed significant deviation from HWE ($P < 0.05$) when the population comprised the 28 and 30 genotypes, respectively (**Figure 3** and **Supplementary Table 4**). Among the 5,531 and 1,687 loci analyzed, 0.5 and 0.2% showed heterozygote deficiency, respectively (**Figure 3**). In total, 28 SNP loci across 27 unigenes exhibited heterozygote deficiency.

Genetic Variation Within and Among Genotypes and Groups

For the genetic diversity analyses, the 5,531 and 1,687 SNPs were used for the two groups comprising 28 and 30 genotypes, respectively (**Table 2**). Among the 5,531 and 1,687 SNP loci, 50 and 37.1% had minor allele frequency (MAF) above 0.05 (**Figures 3C,F**). The analysis using the 5,531 SNPs resulted in observed heterozygosity (Ho) ranging from 0.18 (in genotype Ga01.12) to 0.28 (in genotype Ga08.05), which are the same as the percent polymorphic loci (PPL) of the genotypes. The overall mean observed number of alleles (Na) and Ho were 1.22 and 0.22, respectively. The average genetic distance (GD) of a genotype from the other genotypes ranged from 0.040 (genotype Ga01.20) to 0.055 (genotype Ga10.06), with an overall average of 0.048. Private alleles were detected in 82.1% of the 28 genotypes, with genotype Ga09.03 having the highest number of private alleles (NPA; mean = 0.014). The corresponding analysis using the 1,687 SNPs across the 30 genotypes produced Ho ranging from 0.12 (in genotypes Ga01.12 and Ga01.20) to 0.23 (in genotype

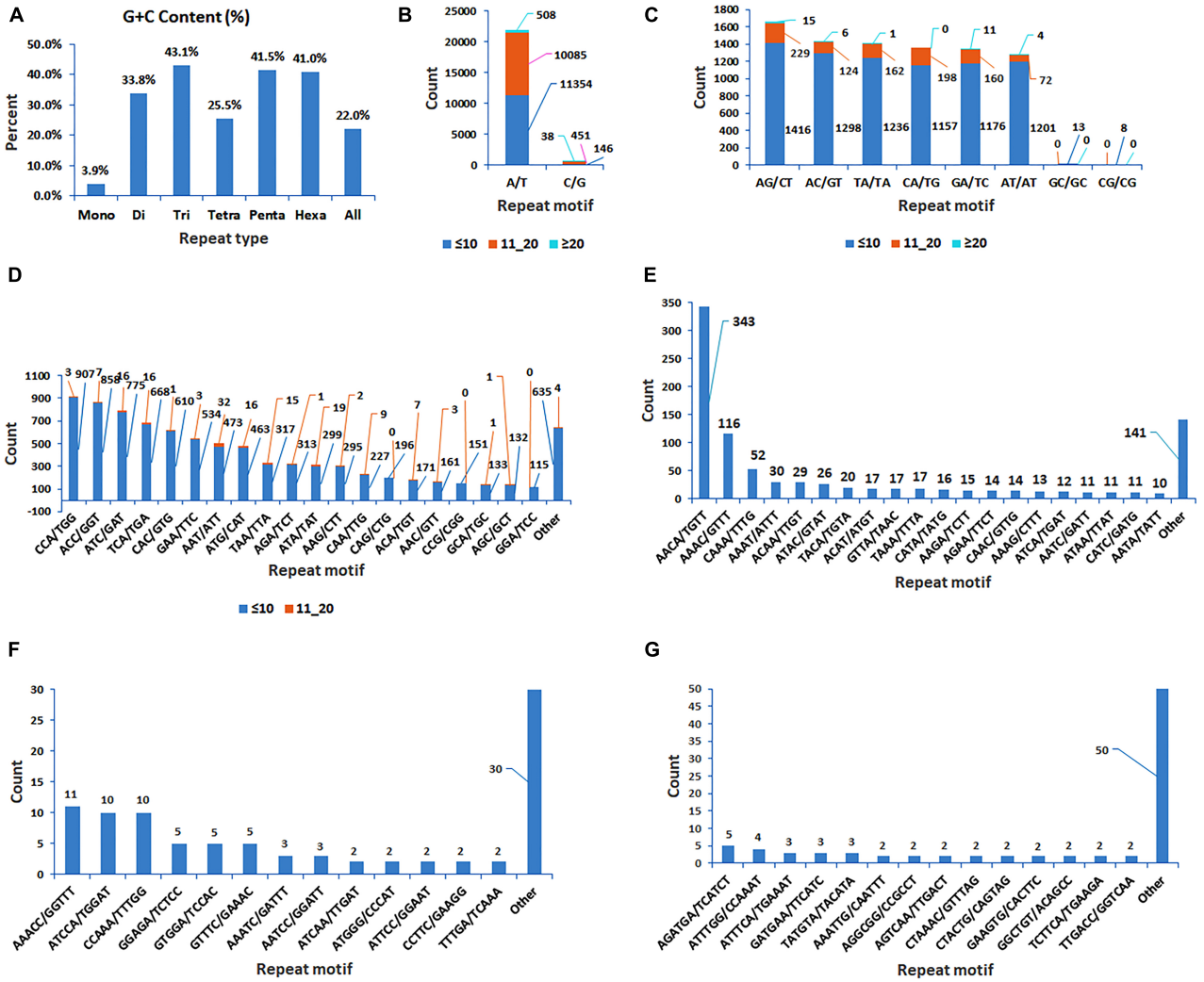


FIGURE 2 | Bar graphs displaying the G+C content of mono, di, tri, tetra, penta and hexanucleotide repeat SSRs, and all SSRs together **(A)**; and the number of classified repeat types considering sequence complementarity) for mono, di, tri, tetra, penta and hexanucleotide repeat SSRs, which were grouped based on the number of times a repeat motif was repeated (three groups for mono and dinucleotide repeats, two groups for trinucleotide repeats, and one group for each of tetra, penta and hexanucleotide repeats) **(B–G)**.

Ga08.05) with an overall mean of 0.18 (**Table 2**). Whereas, an individual genotype's GD from other genotypes ranged from 0.035 (Ga01.12) to 0.051 (Ga10.06), with an overall average of 0.043. In this group, private alleles were detected in all genotypes except in Ga01.12 and Ga01.16 (**Table 2**).

Genotype Ga01.12 and Ga01.08 had a relatively low Nei's distance and mean number of pairwise differences from most of the other genotypes, whereas genotype Ga08.05 had relatively high values in these parameters. The lowest mean number of pairwise differences between genotypes was recorded for Ga01.08 vs Ga01.12 and Ga01.20 vs Ga01.12. The lowest mean number of pairwise differences within genotype was observed in Ga01.12, followed by Ga01.16, Ga01.08, and Ga01.20. In contrast, the mean number of pairwise differences recorded for Ga08.05, Ga08.03, and Ga08.01 was among the highest (**Figure 4**). At a group level, Group-1 had the lowest mean number of pairwise differences within group whereas Group-10 had the highest. Group-1 vs Group-2 had the lowest mean number of pairwise differences between groups, while Group-6 vs Group-10 had the highest Nei's distance (**Figure 4C**).

Cluster, Principal Coordinate and Population Structure Analyses

Neighbor-joining (NJ) cluster analysis and principal coordinate analysis (PCoA) were conducted based on Nei's standard genetic distance (**Supplementary Table 5**) calculated using 5,531 and 1,687 SNP data sets for the 28 and 30 noug genotypes, respectively. Following the approach described in Brown-Guedira et al. (2000) for finding an acceptable number of clusters where the within-cluster genetic distance is below the overall mean genetic distance and where the mean between-cluster

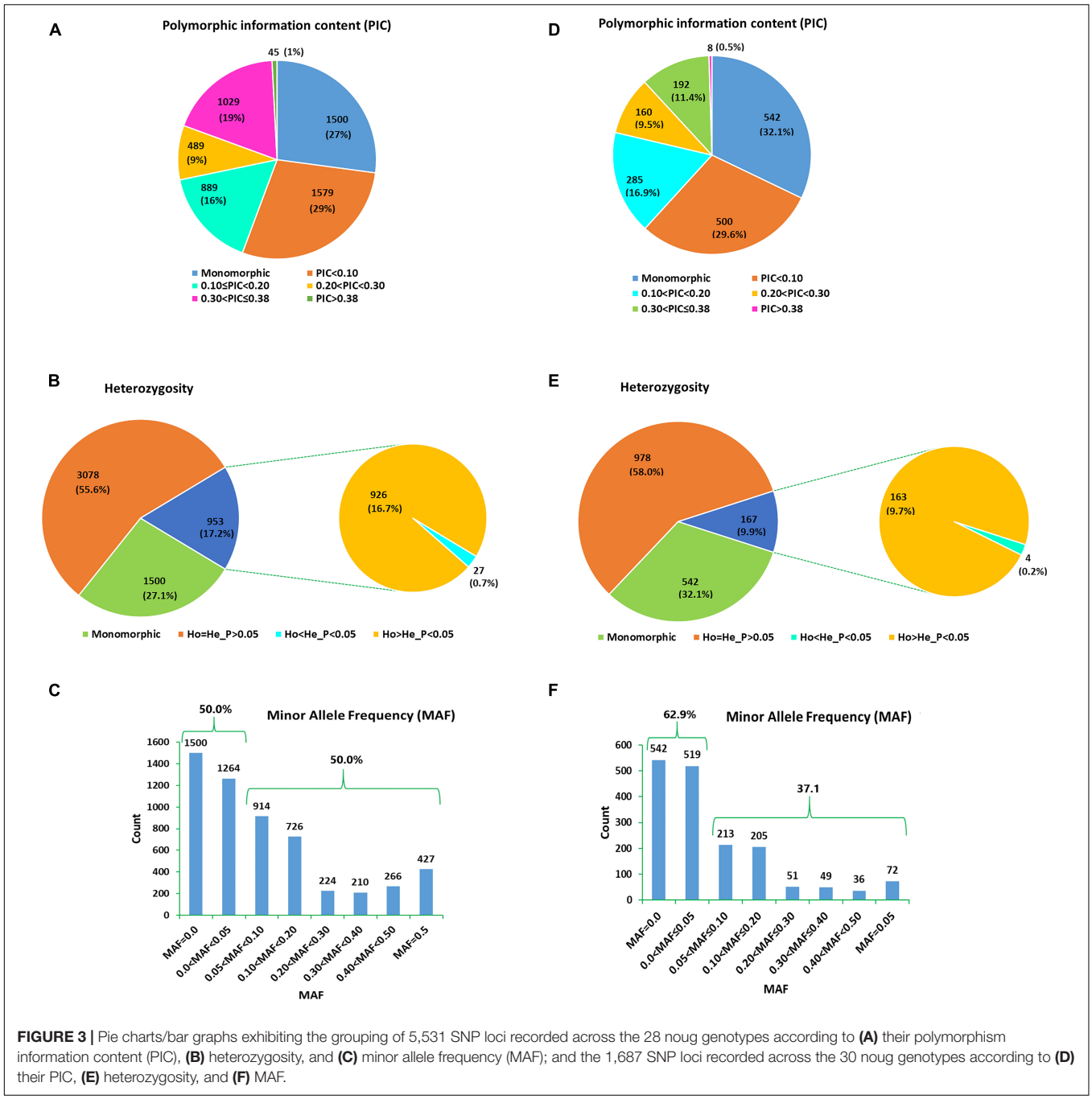


FIGURE 3 | Pie charts/bar graphs exhibiting the grouping of 5,531 SNP loci recorded across the 28 noug genotypes according to **(A)** their polymorphism information content (PIC), **(B)** heterozygosity, and **(C)** minor allele frequency (MAF); and the 1,687 SNP loci recorded across the 30 noug genotypes according to **(D)** their PIC, **(E)** heterozygosity, and **(F)** MAF.

distance is above the mean within-cluster distance, four major clusters were determined in both cases. Even though there are clear differences between the clustering patterns generated with the two data sets, similarities are also evident. In both cases, genotype Ga10.06, Ga08.01, Ga08.05 and Ga09.04, which were assigned to cluster-1 or cluster-2, were among the most differentiated. On the other hand, genotypes that are less sensitive to photoperiod (Ga101B.3 and Ga101B.5) were closely clustered together in cluster-4 (**Figure 5A**) and cluster-2 (**Figure 5B**), respectively. Among the self-compatible genotypes, Ga01-12, Ga01-16 and Ga01-22 (red triangle) were closely clustered in cluster-3 (**Figure 5A**) and Cluster-2 (**Figure 5B**). In both cases, cluster-4 is the most diverse, comprising genotypes from seven of the ten groups (see symbols). In several cases, genotypes within the same phenotypic group were assigned to more than one clusters. For example, both very early-maturing genotypes (blue diamond) and very late-maturing genotypes (red diamond) were placed under more than one cluster (**Figures 6A,B**).

The principal coordinate analysis (PCoA) was conducted to determine the differentiation among the 28 individual genotypes (**Figure 6A**) and the 30 individual genotypes (**Figure 6B**), respectively. In the two-dimensional plots generated, the first

TABLE 2 | Mean values of observed number of alleles (Na), observed heterozygosity (Ho), number of private alleles (NPA), percent polymorphic loci (%PL) for each genotype and Nei's standard genetic distance (GD) of each genotype from all other genotypes based on data from 5,531 SNP loci (for 28 of the 30 genotypes) and 1,687 loci (for all 30 genotypes).

Genotype	28 genotypes and 5,531 Loci					30 genotypes and 1,687 loci				
	Mean Na ± SE	Mean Ho ± SE	Mean NPA ± SE	%PL	GD	Mean Na ± SE	Mean Ho ± SE	Mean NPA ± SE	%PL	GD
Ga01.12	1.18 ± d	0.18 ± d	0.001 ± a	0.18	0.041	1.12 ± f	0.12 ± f	0.000 ± a	0.12	0.035
Ga01.16	1.20 ± d	0.20 ± d	0.001 ± a	0.20	0.041	1.13 ± f	0.13 ± f	0.000 ± a	0.13	0.036
Ga01.22	1.23 ± e	0.23 ± e	0.002 ± b	0.23	0.044	1.15 ± g	0.15 ± g	0.001 ± b	0.15	0.039
Ga01.06	1.23 ± e	0.23 ± e	0.005 ± b	0.23	0.048	1.17 ± g	0.17 ± g	0.004 ± b	0.17	0.044
Ga01.08	1.20 ± d	0.20 ± d	0.001 ± a	0.20	0.045	1.13 ± f	0.13 ± f	0.001 ± b	0.13	0.039
Ga01.20	1.20 ± d	0.20 ± d	0.001 ± a	0.20	0.040	1.12 ± f	0.12 ± f	0.001 ± b	0.12	0.036
Ga02.01	1.23 ± e	0.23 ± e	0.003 ± b	0.23	0.046	1.19 ± g	0.19 ± g	0.003 ± b	0.19	0.042
Ga02.03	1.22 ± e	0.22 ± e	0.004 ± b	0.22	0.045	1.15 ± g	0.15 ± g	0.001 ± b	0.15	0.039
Ga02.07	1.24 ± e	0.24 ± e	0.004 ± b	0.24	0.048	1.17 ± g	0.17 ± g	0.003 ± b	0.17	0.043
Ga01.01	1.26 ± e	0.26 ± e	0.003 ± b	0.26	0.050	1.19 ± g	0.19 ± g	0.003 ± b	0.19	0.045
Ga01.02	1.24 ± e	0.24 ± e	0.004 ± b	0.24	0.048	1.16 ± g	0.16 ± g	0.002 ± b	0.16	0.042
Ga04.11	1.23 ± e	0.23 ± e	0.001 ± a	0.23	0.047	1.16 ± g	0.16 ± g	0.001 ± b	0.16	0.042
Ga02.02	na	na	na	na	na	1.19 ± h	0.19 ± h	0.002 ± b	0.19	0.045
Ga02.06	1.26 ± e	0.26 ± e	0.005 ± b	0.26	0.051	1.18 ± g	0.18 ± g	0.002 ± b	0.18	0.046
Ga04.08	1.24 ± e	0.24 ± e	0.002 ± b	0.24	0.046	1.17 ± g	0.17 ± g	0.002 ± b	0.17	0.040
Ga06.01	1.25 ± e	0.25 ± e	0.008 ± b	0.25	0.050	1.18 ± g	0.18 ± g	0.003 ± b	0.18	0.042
Ga06.02	1.27 ± e	0.27 ± e	0.012 ± b	0.27	0.053	1.21 ± h	0.21 ± h	0.004 ± b	0.21	0.048
Ga09.04	1.25 ± e	0.25 ± e	0.007 ± b	0.25	0.050	1.18 ± g	0.18 ± g	0.004 ± b	0.18	0.045
Ga07.01	1.26 ± e	0.26 ± e	0.005 ± b	0.26	0.051	1.19 ± g	0.19 ± g	0.002 ± b	0.19	0.047
Ga08.01	1.27 ± e	0.27 ± e	0.010 ± b	0.27	0.053	1.21 ± h	0.21 ± h	0.004 ± c	0.21	0.049
Ga09.03	1.26 ± e	0.26 ± e	0.014 ± c	0.26	0.053	1.18 ± g	0.18 ± g	0.008 ± c	0.18	0.046
Ga08.03	1.27 ± e	0.27 ± e	0.008 ± b	0.27	0.052	1.21 ± h	0.21 ± h	0.008 ± c	0.21	0.050
Ga10.02	1.25 ± e	0.25 ± e	0.000 ± a	0.25	0.049	1.18 ± g	0.18 ± g	0.004 ± b	0.18	0.042
Ga10.06	1.27 ± e	0.27 ± e	0.000 ± a	0.27	0.055	1.20 ± h	0.20 ± h	0.007 ± c	0.20	0.051
Ga08.05	1.28 ± e	0.28 ± e	0.007 ± b	0.28	0.054	1.23 ± h	0.23 ± h	0.003 ± b	0.23	0.049
Ga09.02	1.25 ± e	0.25 ± e	0.005 ± b	0.25	0.048	1.17 ± g	0.17 ± g	0.002 ± b	0.17	0.042
Ga10.08	1.24 ± e	0.24 ± e	0.000 ± a	0.24	0.046	1.18 ± g	0.18 ± g	0.001 ± b	0.18	0.042
Ga101B.3	1.25 ± e	0.25 ± e	0.000 ± a	0.25	0.050	1.18 ± g	0.18 ± g	0.002 ± b	0.18	0.045
Ga101B.5	1.26 ± e	0.26 ± e	0.000 ± a	0.26	0.052	1.19 ± h	0.19 ± h	0.001 ± b	0.19	0.046
Ga101B.m	na	na	na	na	na	1.18 ± g	0.18 ± g	0.002 ± b	0.18	0.044
Mean	1.22 ± b	0.22 ± b	0.004 ± b	0.24	0.048	1.19 ± c	0.18 ± c	0.003 ± b	0.17	0.043

± SE = standard error with a, b, c, d, e, f, g, and h equal to 0, 0.001, 0.002, 0.005, 0.006, 0.008, 0.009, and 0.01; respectively.
na = Not applicable.
The Pearson correlation coefficient between the two groups for NA, Ho and %PL was 0.94 ($P < 0.001$); for NPA was 0.59 ($P = 0.001$), and for GD was 0.95 ($P < 0.001$).

and the second coordinates explained 10.2 and 7.3% of the total variation among the 28 individual genotypes (**Figure 6A**) and 7.1 and 6.1% of the total variation among the 30 individual genotypes (**Figure 6B**), respectively. Hence, the two coordinates together explained 17.5% of the total variation among the 28 individual genotypes and 13.2% of the total variation among the 30 individual genotypes, both of which are quite low. However, the clustering pattern of the genotypes in both two-dimensional plots (**Figures 6A,B**) are in good agreement, as clusters highlighted by the same color mostly represent the same genotypes. Most self-compatible genotypes (see **Supplementary Table 1**) were assigned to the light-blue highlighted clusters. The results of PCoA and cluster analysis also agree well in general. For example, similar to cluster analysis, genotypes less sensitive to photoperiod were closely clustered in PCoA (pink highlighted genotypes in **Figure 6B**). Analyses of the population genetic structure based on admixture models using 5,531 SNPs for the 28 genotypes and 1,687 SNPs for the 30 genotypes demonstrated that the genotypes are best represented by three genetic populations. ($K = 3$; **Supplementary Figures 1A,C**). It is interesting to note that each genotype has alleles originating from the three genetic populations, in both cases, demonstrating a strong genetic admixture (**Supplementary Figures 1B,D**).

DISCUSSION

The SSR Characteristics in Noug Unigenes

The RNA-Seq based sequencing of 30 noug genotypes resulted in 409,309 unigenes ranging in size from 201 to 13,568 bp, totaling 204.2 Mbp, and having a G+C content of 40%. The G+C content is an important feature of genome organization, and show wide variation among different genomes and different regions within a genome, and has been studied in connection with understanding genome evolution (Morgante et al., 2002;

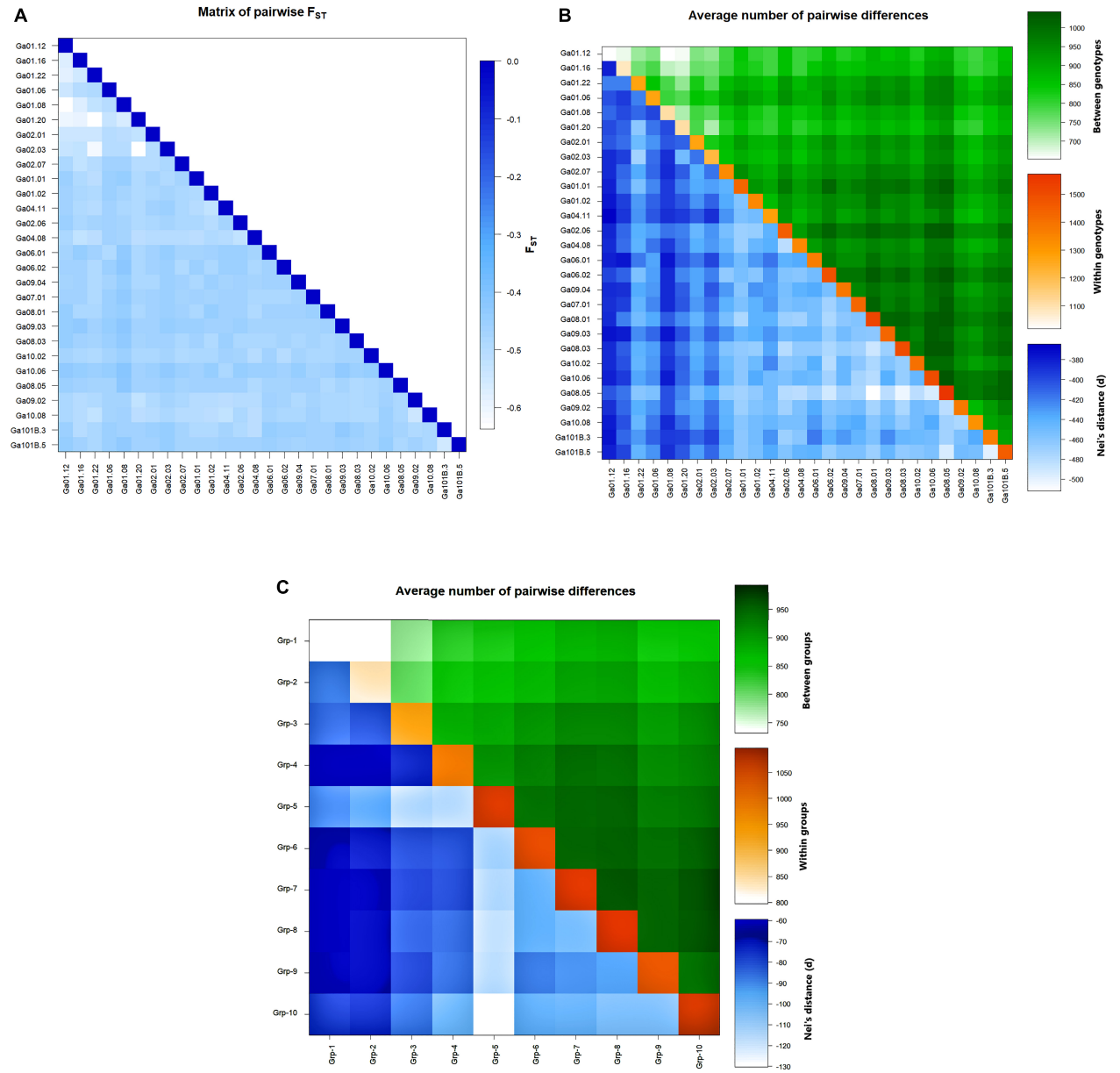


FIGURE 4 | Heatmaps of **(A)** pairwise F_{ST} between the 28 genotypes, **(B)** average number of pairwise differences (above diagonal) and Nei's distance (below diagonal) between the 28 genotypes, and average number of pairwise differences within the 28 genotypes (diagonal), and **(C)** average number of pairwise differences (above diagonal) and Nei's distance (below diagonal) between the 10 groups of genotypes, and average number of pairwise differences within the 10 group of genotypes (diagonal).

Glémin et al., 2014; Singh et al., 2016). Diversity in G+C content in plant genomes is biologically and evolutionary significant, including its importance for plant adaptation to diverse environments (Šmarda et al., 2014). Studies have shown that grasses, such as rice and maize, have genomic G+C content above 40%, while dicots have G+C content below 40% (Wang et al., 2004; Qin et al., 2015; Singh et al., 2016). In general, genes have a higher G+C content than genomic sequences, with their coding sequences (CDS) having a higher G+C content than their 3′ and 5′ untranslated regions (3′-UTR and 5′-UTR) (Zhao et al., 2014; Singh et al., 2016). The G+C content of CDS exceeds 40% even for dicots (Wang et al., 2004; Singh et al., 2016). Hence, the G+C content of 40% obtained in the present study for the noug unigenes (CDS plus UTRs) is consistent with data reported for other dicots.

Simple sequence repeats (SSRs) are ubiquitous and highly polymorphic loci in plant genomes comprising tandemly repeated nucleotide sequences of 1 to 6 bp in length. Genomic events that lead to the length polymorphism of SSRs include unequal recombination between homologous SSRs and

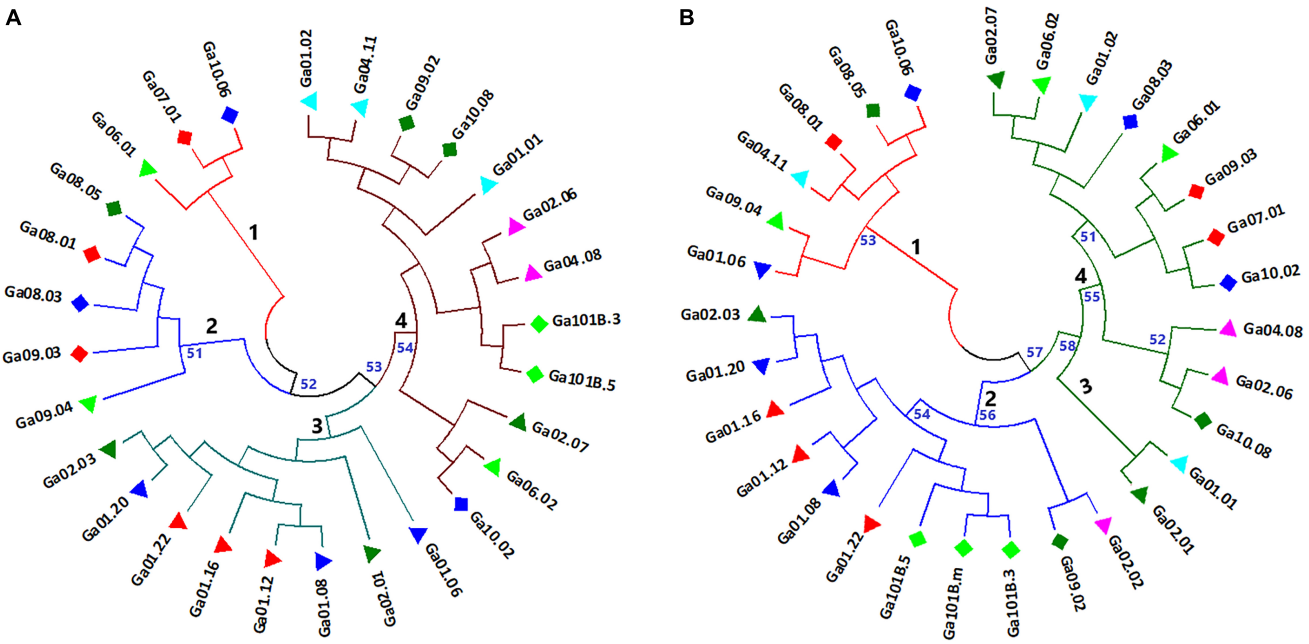


FIGURE 5 | Neighbor-joining trees depicting the clustering pattern of **(A)** 28 noug genotypes and **(B)** 30 noug genotypes based on Nei's standard genetic distances calculated using genotypic data at 5,531 and 1,687 SNP loci, respectively. The numbers in black represent the four clusters of each tree, and the numbers in blue represent the bootstrap support of the corresponding branches (only bootstrap values above 50 are shown).

replication slippage that result in repeat motif deletion or insertion (Li et al., 2004). In the CDS, frameshift mutations that result in a gain or loss of function occur as the result of insertions or deletions of the SSR repeat motifs (Li et al., 2004). The high mutation rate of SSRs makes them a significant component of genome evolution (Kashi et al., 1997), and they are excellent molecular markers for various applications (Olmstead et al., 2008; Geleta et al., 2012; Lu et al., 2012; Shiferaw et al., 2012; Shirasawa et al., 2013; Teshome et al., 2015; Chombe et al., 2017). The distribution and density of SSRs vary among genomes of different species as well as different regions within genomes (Tóth et al., 2000; Temnykh et al., 2001; Mun et al., 2006). Similarly, the frequency of different types of SSRs (mono, di, tri, tetra, penta, and hexanucleotide repeats) as well as the nucleotide composition of their repeat motifs differ within and among genomes (Morgante et al., 2002; Grover and Sharma, 2007; Qin et al., 2015).

Mononucleotide repeats were the most frequent in the present study accounting for over half of the SSRs identified, of which A/T SSRs accounted for 97.2%. Similarly, AT/TA SSRs were by far more prevalent among dinucleotide repeat SSRs than CG/GC SSRs, accounting for 32.4 and 0.24% of all dinucleotide repeat SSRs, respectively. Such an overwhelming dominance of A/T over C/G and AT/TA over CG/GC in noug unigenes is consistent with that of mono and dinucleotide repeat SSRs in the genomes of other dicots, including *Arabidopsis thaliana*, *Glycine max*, *Vitis vinifera* and *Solanum lycopersicum* (Qin et al., 2015). According to Qin et al. (2015), C/G and CG/GC SSRs declined during the evolution of plant genomes, which warrants further research to identify the major causes for this change. In other groups of dinucleotide repeat SSRs, homopurine/homoprymidine motifs (AG/CT+GA/TC) were more frequent than purine-prymidine mix (AC/GT+CA/TG) in the present study, in agreement with other studies in dicots (Grover and Sharma, 2007; Wang et al., 2008; Qin et al., 2015). The relative frequency of tri, tetra, penta, and hexanucleotide SSR motifs differed among studies, even within dicots, in contrast to mono and dinucleotide repeat SSRs. Trinucleotide repeat SSRs are more common than dinucleotide repeat SSRs in Arabidopsis CDS and UTRs (Morgante et al., 2002), unlike the case in the present study, where they are more or less equally frequent. The two most common trinucleotide SSRs in the present study were those with ACC/GGT and CCA/TGG motifs, unlike in Papaya where the AAG motif dominates the trinucleotide SSRs (Wang et al., 2008). Among complimentary motifs, notable differences exist between GGT and ACC, and between GAA and TTC, accounting for 6.4, 3.7, 4.0, and 2.2% of trinucleotide repeat SSRs, respectively. Hence, higher frequencies of GGT and GAA in the transcribed sequences of noug require further research in comparison with other dicots.

The G+C content in the noug unigenes (40%) is significantly higher than the G+C content of the SSRs derived from these unigenes (22.2%). Similar pattern was reported in *Populus* where 33.2% G+C content in the whole genome but 25.4% in the SSRs (ShuXian and TongMing, 2007). The CCG/CGG trinucleotide repeats are abundant in monocots (rice, maize, and wheat) but rare in dicots (Arabidopsis and soybean) (Morgante et al., 2002). They are among low-frequency trinucleotide SSRs in the present study, which is similar to the results from Morgante et al. (2002) study for dicots. Also, they found higher G+C content in monocot ESTs than in dicot ESTs. Nevertheless,

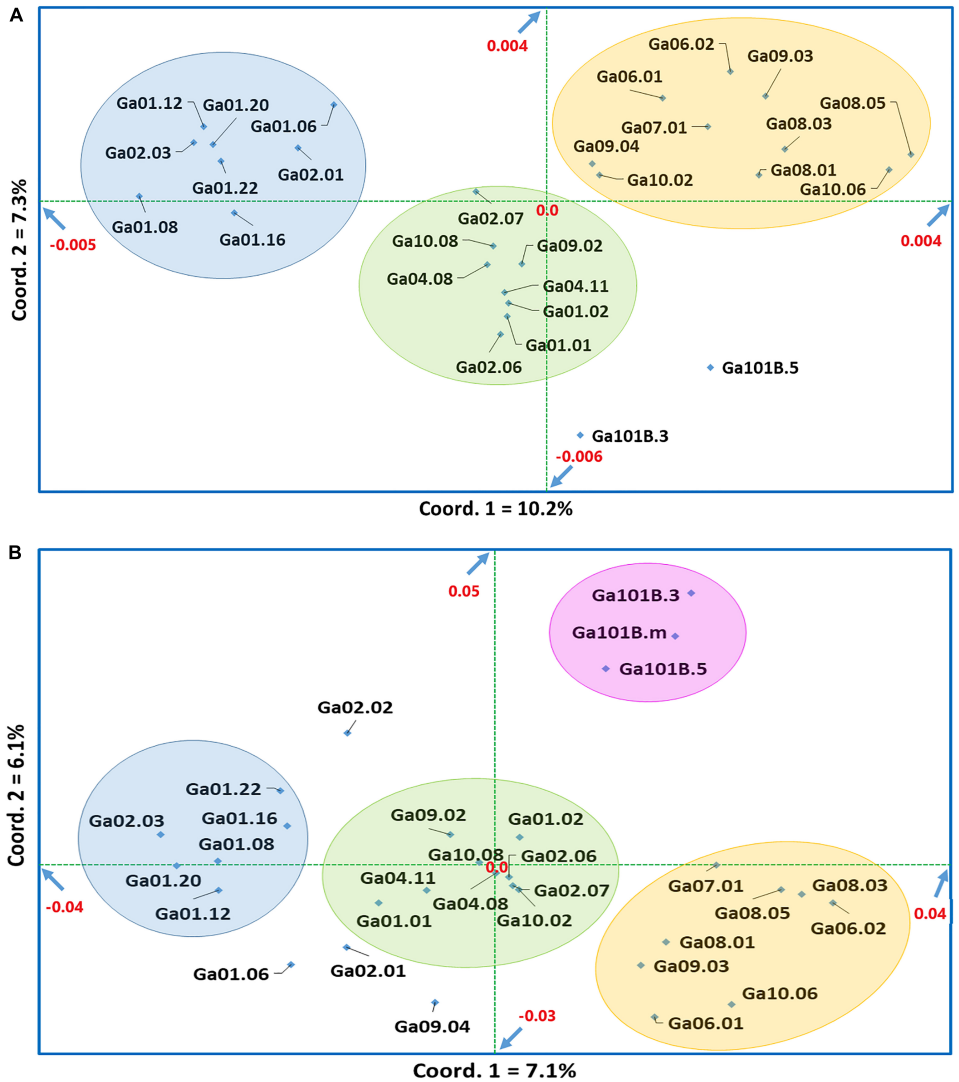


FIGURE 6 | Principal coordinate analysis (PCoA) of **(A)** 28 noug genotypes and **(B)** 30 noug genotypes based on Nei's standard genetic distances calculated using genotypic data at 5,531 and 1,687 SNP loci, respectively. The first two principal coordinates together explained 17.5 and 13.2% of the total variation in the case of panels **(A,B)**, respectively.

the G+C content of 44% in EST-derived Arabidopsis and soybean SSRs they reported is twice that of the present study's noug SSRs (22%). The marked difference between the two studies could be partly attributed to differences in the representation of CDS and UTRs in the respective sequences; further studies will shed more light on this. The higher G+C content of trinucleotide SSRs than di and tetraploid SSRs in the present study is most likely due to the greater number of GC-rich trinucleotide SSRs in CDS, which do not cause frameshift mutations.

Using transcriptome-based SSR markers, previous studies on noug revealed higher genetic variation both within and between populations (Dempewolf et al., 2010, 2015; Misganaw and Abera, 2017) in comparison to results obtained using dominant markers (Geleta et al., 2007, 2008) and bi-allelic SNP markers (Tsehay et al., 2020), indicating the superiority of multi-allelic SSR markers in their informativeness. Since a reference genome sequence is not available for noug yet, the genomic positions of the SSRs identified in the present study is currently unknown. With the annotation of the unigenes used, it will be possible to select genome-wide single-copy SSRs for genotyping a diverse panel of noug genotypes and then develop gene-associated polymorphic SSRs for their numerous applications in noug and other *Guizotia* species, as the rate of cross-species transferability of transcriptome-derived SSR markers is proved high (Lu et al., 2013; Teshome et al., 2015; Zhou et al., 2016; Gadissa et al., 2018; Serbessa Tolera et al., 2021). The applications include genetic diversity analyses for conservation and breeding, as well as genetic linkage mapping and genome-wide association studies.

The SNP Markers and Genetic Variation Among Genotypes

Given that noug is strictly outcrossing in general (Nemomissa et al., 1999; Geleta and Bryngelsson, 2010; Geleta and Ortiz, 2013), observed heterozygosity (Ho) is expected to equal or exceed the expected heterozygosity (He) if other HWE assumptions are met. However, a very small fraction of the SNP loci (< 0.5%) exhibited heterozygote deficiency. Hence, natural selection may be favoring homozygosity at these loci although self-pollination might have contributed to the heterozygote deficiency given that 40% of the genotypes are self-compatible to different extents. Contrary to this, 9.9% of the loci (**Supplementary Table 4**) were heterozygous across all genotypes, which is particularly interesting when considering loci with proportional allele frequencies. Natural selection favoring heterozygosity might have contributed, but genotype calling based on reads from duplicate genes with different alleles cannot be ruled out. The development of a reference genome sequence for noug, as well as the annotation and comparison of the unigenes with sunflower genes (the closest to noug among well-studied crops) will provide evidence that explain these results.

As the number of markers increased from 1,687 (Ho = 0.18) to 5,531 (Ho = 0.22), the mean observed heterozygosity (Ho) also increased, suggesting that the number of markers influences the parameter, particularly for small number of samples. Whereas, a study on 24 noug accessions comprising 281 genotypes reported a slightly higher Ho (0.24) based on 202 transcriptome derived-polymorphic SNP markers (Tsehay et al., 2020), suggesting a stronger effect of sample size than number of markers. Considering the analysis of 28 genotypes using 5,531 markers, Ho varied from 0.18 (Ga01.12) to 0.28 (Ga08.05). On average, self-compatible genotypes were less heterozygous than their self-incompatible counterparts were, and the lower Ho values in self-compatible genotypes resulted from self-pollination. There is, however, still substantial heterozygosity in self-compatible genotypes that have been self-pollinated for a number of generations. As inbreeding depression in noug is high (Geleta and Bryngelsson, 2010; Geleta and Ortiz, 2013), a significant proportion of plants grow poorly following self-pollination. As a result, selecting plants with higher proportions of heterozygous loci for next round breeding is more likely, explaining the high heterozygosity of the self-compatible genotypes. Consequently, developing pure-line cultivars is likely to be challenging although self-compatible genotypes were successfully developed.

Polymorphism information content (PIC) measures the usefulness of DNA markers in terms of detecting genetic variation (Hildebrand et al., 1992; Shete et al., 2000). The PIC of a locus depends on the number and frequency of its alleles, which in turn depends on the diversity of genotypes (populations) analyzed. In the larger data set 5,531 SNPs, 19.4% had a PIC of above 0.30, which makes them highly informative. In a similar study in noug, 50% of the 202 markers used had a PIC value above 0.25 (Tsehay et al., 2020). Comparatively, 31% of polymorphic markers had PIC values above 0.25 in this study. The lower proportion in this study can be explained by a smaller number of samples used compared to Tsehay et al. (2020). Nevertheless, 1,266 SNP markers had PIC exceeding 0.25 (1,074 of which had PIC above 0.30; **Supplementary Table 3**), which can be prioritized for use in various applications, including population genetics for conservation and breeding, genetic linkage mapping, and genome-wide association studies.

There is a good correlation between results obtained from the analyses of the data sets containing 5,531 and 1,687 SNPs, although the values of most parameters analyzed are higher for the larger data set. In both cases, the highest mean genetic distance was recorded in genotype Ga10.06 and the highest number of private alleles was recorded in genotype Ga09.03. Both genotypes are self-incompatible but they mature at different times. The genotype Ga10.06 was sampled from a very early-type landrace population that was originally collected from Arsi (39 km from Bekoji to Tereta; southeast Ethiopia), whereas the genotype Ga09.03 was sampled from a very late-type population that was originally collected from Gojjam (35 km from Amanuel to Bure; northwest Ethiopia). A higher mean genetic distance of Ga10.06 is not surprising since it came from an isolated location where the cultivation of noug is low. Ga09.03 was sampled from a major noug growing region that it shared with the other two very late-type genotypes (Ga07.01 and Ga08.01), so its relatively high number of private alleles was noteworthy given the high rate of gene flow within the region (Geleta et al., 2008).

The lowest mean number of pairwise differences (MNPD) among genotypes were recorded between pairs of self-compatible genotypes (Ga01.08 vs Ga01.12 and Ga01.20 vs Ga01.12). Self-compatible genotypes are developed through crossbreeding of a few genotypes that exhibit a low level of self-compatibility, and hence, their low pairwise differences is due to their narrow genetic basis. The lowest mean number of pairwise differences within genotypes (e.g., heterozygosity) was also recorded in self-compatible genotypes, which is not surprising since the genotypes have been self-pollinated for a number of generations, and hence increased homozygosity as compared to the self-incompatible genotypes. Those that exhibited the highest mean number of pairwise differences within genotypes (Ga08.05, Ga08.03, and Ga08.01) are all strictly self-incompatible.

Genetic Variation of Genotypes Within Trait-Based Groups

The 30 noug genotypes used in the present study were grouped into 10 different groups based on their phenotypic characteristics. Each group differs from the others at least in one characteristic in terms of ability to set self-seeds, photoperiod sensitivity, duration to reach seed maturity, and seed oil and oleic acid contents. However, the genotypes within each group were genetically diverse with the exception of Group-1 (Ga01.12, Ga01.16, and Ga01.22) comprising genotypes bred for higher oil content, and Group-10 (Ga101B.3, Ga101B.5, and Ga101B.m) comprising genotypes with a lower photoperiod sensitivity (**Supplementary Table 1**).

Overall, the self-compatible groups were more closely related to one another than the self-incompatible ones. The self-compatible genotypes were developed through crossbreeding and selfing based on a limited number of genotypes originating

from a few landrace populations. As such, their relatively higher genetic relationship is a result of their narrow genetic base and the crossbreeding scheme used. Interestingly, both the cluster analysis and principal coordinate analysis assigned genotypes with oil content above 40% to more than one cluster. For example, Ga01.20 and Ga02.07 are both high oil content genotypes (over 40%) and self-compatible genotypes (Ga01.20 being among the best for self-compatibility) but they were assigned to different clusters in both analyses. Hence through crossbreeding these genotypes, a self and cross-pollinating cultivar with high seed and oil yields can be developed. It would be very interesting to apply such an approach to noug, as it can overcome the potential consequences of inbreeding depression.

The dominant fatty acid in noug seed oil is linoleic acid (C18:2) and oleic acid (C18:1) content is generally below 13%, particularly in noug grown in Ethiopia (Dagne and Johnson, 1997; Geleta et al., 2011; Tsehay et al., 2020). However, genotypes with C18:1 above 13% have been identified and crossbred to develop high oleic acid types (Geleta et al., 2011; Geleta and Ortiz, 2013) although their oleic acid levels fluctuate with the average temperature of the growing environments. They produce significantly higher C18:1 at the expense of C18:2 in low-altitudes [below 1,800 meters above sea level (masl)] than in high-altitudes (above 2,200 masl) (Geleta et al., 2011; Tsehay et al., 2020). Among the genotypes included in the present study, three self-compatible genotypes (Ga01.16, Ga02.01, and Ga01.02) and one self-incompatible genotype (Ga02.02) had an oleic acid content above 13%, except when grown in high-altitude environments. The data analyses revealed that these genotypes are genetically diverse and differ in desirable traits, such as oil content. Therefore, their crossbreeding may result in high-oleic acid noug cultivars suitable for low-altitude cultivation. Early maturity is a highly desirable trait in crops, especially when the growing season is short or in drought-prone areas, but it usually comes at a cost in terms of yield (Cattivelli et al., 2008). The genotypes included, in the present study varied from "very-early" type to "very-late" type, which took ca 120 and 180 days from planting to harvesting, respectively, when grown at a high-altitude location (Holeta agricultural research center in Ethiopia; 9°00′ N, 38°30′ E; 2400 masl). Based on pairwise comparison as well as cluster and principal coordinate analyses, Group-8 (Ga08.03, Ga10.02, and Ga10.06) consisted of very-early type self-incompatible genotypes, which are genetically diverse. Crossbreeding these genotypes can therefore improve various desirable traits without affecting their earliness.

Research in population genetics uses various approaches to determine the genetic structure of populations and the source of genotypes (Rannala and Mountain, 1997; Davies et al., 1999; Pritchard et al., 2000; Alexander et al., 2009; Raj et al., 2014). In the present study, a model-based approach of Pritchard et al. (2000) was used for population structure analysis, which assumes that populations are characterized by a set of allele frequencies across multiple loci. By using this approach, each individual within a predefined population is probabilistically assigned to a cluster, or it is assigned to multiple clusters if it is determined to be admixed. The genotypes in the present study were analyzed to determine the population genetic structure using this model. The analysis using the Puechmaille (2016) approach determined that the optimal number of clusters (K) is three, corresponding to three genetic populations. Interestingly, all genotypes are the results of admixture from the three genetic populations with a slightly different extent. This significant level of admixture may have caused the discrepancy between the four clusters obtained from cluster analysis and PCoA compared to the three clusters obtained from Bayesian statistics-based population genetic structure analysis. A recent study on 24 diverse noug accessions comprising 281 genotypes also revealed three genetic populations with strong admixture (Tsehay et al., 2020). The studies generally suggest a weak population structure in noug due to population admixture caused by strong gene flow between populations *via* pollen and germplasm exchange that gradually covers wide geographic areas.

CONCLUSION

Through RNA-Seq based sequencing, 409,309 unigenes, representing the noug transcriptome, have been developed for its various applications in the present study. The G+C content of these unigenes was 40%, which is comparable to that of other dicots. The analyses of SSRs in the unigenes revealed an overwhelming predominance of A/T over C/G and AT/TA over CG/GC, consistent with other dicots. Interestingly, GGT and GAA repeats had a higher frequency than their complementary motifs. This suggests their greater importance in noug genes, and therefore requires further investigation in comparison with other dicots. The whole unigenes are significantly higher in G+C content (40%) than the SSRs derived from them (22.2%). Further research and analysis of the SSRs identified in the current study could lead to the development of genome-wide single-copy SSRs with high polymorphism for use in noug breeding and research. Thousands of high-quality SNPs were discovered in each noug genotype in the present study, and well over a thousand of them were common to all genotypes and possessed a high polymorphism information content (PIC $>$ 0.30), which makes them ideal for use in a wide range of applications. The significant levels of admixture observed in each noug genotypes suggest a weak population structure in noug likely caused by strong gene flow between populations across wide geographic areas. Although the self-compatible genotypes were bred for several generations with self-pollination, a substantial level of heterozygosity was observed, suggesting an inbreeding depression that led to plants with higher heterozygosity being selected in successive generations, presenting potential challenges to the development of highly productive and nutritionally rich pure-line cultivars. Interestingly, genotypes that share desirable characteristics, such as self-compatibility, early maturity, high oil content, or high oleic acid content are genetically diverse. Crossbreeding these genotypes would enable the development of cultivars that combine these characteristics and reproduce through both selfing and cross-pollination, which would be a viable approach to overcome the potential effects of inbreeding depression.

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: National Center for Biotechnology Information (NCBI) BioProject database under accession numbers GJSF00000000 and PRJNA763316 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA763316).

AUTHOR CONTRIBUTIONS

MG and RO: conceptualization. AG, MG, RV, and CH: methodology. AG and MG: software. AG, MG, CH, and RO: data analysis. AG: writing–original draft. RO, MG, and KT: funding acquisition. All authors have contributed in supervision, writing—review and editing, read and agreed to the published version of the manuscript.

FUNDING

This study was financed by the Swedish International Development Cooperation Agency (Sida) through the research and training grant awarded to Addis Ababa University and the Swedish University of Agricultural Sciences (AAU-SLU Biotech; https://sida.aau.edu.et/index.php/biotechnology-phd-program/, and the Swedish Research Council (Vetenskapsrådet, VR) through the collaborative development research project 2014-03517 between SLU, AAU and the Ethiopian Institute of Agricultural Research (EIAR).

ACKNOWLEDGMENTS

We thank the Swedish International Development Cooperation Agency (Sida) and the Swedish Research Council (Vetenskapsrådet, VR) for financing this research. We would also like to thank the Institute of Biotechnology, Addis Ababa University and Department of Plant Breeding, Swedish University of Agricultural Sciences, for technical support during the course of the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.882136/full#supplementary-material

Supplementary Figure 1 | Graphs depicting three clusters (K = 3) as the best representation of **(A)** the 28 noug genotypes based on 5,531 SNP loci, and **(C)** the 30 genotypes based on 1,687 SNP loci, using the method of Puechmaille (2016) for the determination of the optimum number of clusters; and the corresponding graphical display of the genetic structure of **(B)** 28 noug genotypes and **(D)** 30 noug genotypes generated based on genotypic data at 5,531 and 1,687 SNP loci, respectively, following the determination of the optimum number of clusters (K) of three (K = 3) using the method. The three colors in **(B)** and **(C)** correspond to the three clusters (genetic populations) and the proportion of each color in each genotype denotes the average proportion of the alleles that placed each accession under the three clusters.

Supplementary Table 1 | Plant material (genotypes) used for this study and their general description.

Supplementary Table 2 | The list of SSRs detected within the 35639 Unigenes and their descriptions.

Supplementary Table 3 | List of the 5531 SNP loci that passed the quality filtering criteria and recorded accross 28 noug genotypes, together with their SNP position (POS), reference allele (REF), alternative allele(s) (ALT), quality score, (QUAL), polymorphic information content (PIC), Observed heterozygosity (Ho), expected heterozygosity (He), P-value for Hardy-Weinbeg Equlaibrium (HWE P-value) and corrsponding reference unigene sequence.

Supplementary Table 4 | List of SNP loci that showed signfcant deviation from Hardy-Weinberg Equlibrium (HWE), under the assumption that the 28 genotypes represent a single population, together with their SNP position (POS), reference allele (REF), alternative allele(s) (ALT), Observed heterozygosity (Ho), expected heterozygosity (He), Ho-He and P-value for Hardy-Weinbeg Equlaibrium (HWE P-value).

Supplementary Table 5 | The pairwise Nei's standard genetic distance between the 28 noug genotypes calculated based on 5531 SNP loci. The diagonal values are mean genetic distance of each genotype from all other genotypes. The Pearson correlation coefficent between the mean genetic distance for 28 genotypes (5531 polymorphic loci) and the 30 genotypes (1687 polymorphic loci) was 0.955 ($P < 0.00001$).

REFERENCES

Alemaw, G., and Alamayehu, N. (1997). *Highland Oilcrops: A Two-Decade Research Experience in Ethiopia. In Research Report No. 30.* Addis Ababa: Institute of Agricultural Research.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi: 10.1093/bioinformatics/btx198

Brown-Guedira, G. L., Thompson, J. A., Nelson, R. L., and Warburton, M. L. (2000). Evaluation of genetic diversity of soybean introductions and North American ancestors using RAPD and SSR markers. *Crop Sci.* 40, 815–823. doi: 10.2135/cropsci2000.403815x

Cattivelli, L., Rizza, F., Badeck, F.-W., Mazzucotelli, E., Mastrangelo, A. M., Francia, E., et al. (2008). Drought tolerance improvement in crop plants: an integrated view from breeding to genomics. *Field Crops Res.* 105, 1–14. doi: 10.1016/j.fcr.2007.07.004

Chombe, D., Bekele, E., Bryngelsson, T., Teshome, A., and Geleta, M. (2017). Genetic structure and relationships within and between cultivated and wild korarima [*Aframomum corrorima* (Braun) PCM Jansen] in Ethiopia as revealed by simple sequence repeat (SSR) markers. *BMC Genet.* 18:72. doi: 10.1186/s12863-017-0540-4

Cloonan, N., Forrest, A., Kolle, G., Gardiner, B. A., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619. doi: 10.1038/nmeth.1223

Dagne, K. (1994). Meiosis in interspecific hybrids and genomic interrelationships in *Guizotia* Cass. (Compositae). *Hereditas* 121, 119–129. doi: 10.1111/j.1601-5223.1994.00119.x

Dagne, K., and Johnson, A. (1997). Oil content and fatty acid composition of seeds of *Guizotia abyssinica* (L.f.) Cass (Compositae). *J. Sci. Food Agric.* 73, 274–278. doi: doi: 10.1002/(sici)1097-0010(199703)73:3<274::aid-jsfa725>3.0.co;2-f

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *Gigascience* 10:giab008. doi: 10.1093/gigascience

Davies, N., Villablanca, F. X., and Roderick, G. K. (1999). Determining the source of individuals: multilocus genotyping in nonequilibrium population

genetics. *Trends Ecol. Evol.* 14, 17–21. doi: 10.1016/s0169-5347(98)01530-4

Dempewolf, H., Kane, N. C., Ostevik, K. L., Geleta, M., Barker, M. S., Lai, Z., et al. (2010). Establishing genomic tools and resources for *Guizotia abyssinica* (Lf) Cass.—the development of a library of expressed sequence tags, microsatellite loci, and the sequencing of its chloroplast genome. *Mol. Ecol. Resour.* 10, 1048–1058. doi: 10.1111/j.1755-0998.2010.02859.x

Dempewolf, H., Tesfaye, M., Teshome, A., Bjorkman, A. D., Andrew, R. L., Scascitelli, M., et al. (2015). Patterns of domestication in the Ethiopian oil—seed crop noug (*Guizotia abyssinica*). *Evol. Appl.* 8, 464–475. doi: 10.1111/eva.12256

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498. doi: 10.1038/ng.806

Ethiopian Institute of Agricultural Research [EIAR] (2017). *Oilseed Crops Strategy 2016-2023*. Addis Ababa: EIAR.

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Gadissa, F., Tesfaye, K., Dagne, K., and Geleta, M. (2018). Genetic diversity and population structure analyses of Plectranthus edulis (Vatke) Agnew collections from diverse agro-ecologies in Ethiopia using newly developed EST-SSRs marker system. *BMC Genet.* 19:92. doi: 10.1186/s12863-018-0682-z

Gebeyehu, A., Hammenhag, C., Ortiz, R., Tesfaye, K., and Geleta, M. (2021). Characterization of Oilseed Crop Noug (*Guizotia abyssinica*) using agro-morphological traits. *Agronomy* 11:1479. doi: 10.3390/agronomy11081479

Geleta, M. (2007). *Genetic Diversity, Phylogenetics and Molecular Systematics of Guizotia Cass.(Asteraceae)*. Uppsala: Swedish University of Agricultural Sciences.

Geleta, M., Asfaw, Z., Bekele, E., and Teshome, A. (2002). Edible oil crops and their integration with the major cereals in North Shewa and South Welo, Central Highlands of Ethiopia: an ethnobotanical perspective. *Hereditas* 137, 29–40. doi: 10.1034/j.1601-5223.2002.1370105.x

Geleta, M., and Bryngelsson, T. (2010). Population genetics of self-incompatibility and developing self-compatible genotypes in niger (*Guizotia abyssinica*). *Euphytica* 176, 417–430. doi: 10.1007/s10681-010-0184-1

Geleta, M., Bryngelsson, T., Bekele, E., and Dagne, K. (2007). Genetic diversity of *Guizotia abyssinica* (L. f.) Cass.(Asteraceae) from Ethiopia as revealed by random amplified polymorphic DNA (RAPD). *Genet. Resour. Crop Evol.* 54, 601–614. doi: 10.1007/s10722-006-0018-0

Geleta, M., Bryngelsson, T., Bekele, E., and Dagne, K. (2008). Assessment of genetic diversity of *Guizotia abyssinica* (Lf) Cass.(Asteraceae) from Ethiopia using amplified fragment length polymorphism. *Plant Genet. Resour* 6, 41–51. doi: 10.1017/s1479262108913903

Geleta, M., Heneen, W. K., Stoute, A. I., Muttucumaru, N., Scott, R. J., King, G. J., et al. (2012). Assigning Brassica microsatellite markers to the nine C-genome chromosomes using Brassica rapa var. trilocularis–B. oleracea var. alboglabra monosomic alien addition lines. *Theor. Appl. Genet.* 125, 455–466. doi: 10.1007/s00122-012-1845-3

Geleta, M., and Ortiz, R. (2013). The importance of *Guizotia abyssinica* (niger) for sustainable food security in Ethiopia. *Genet. Resour. Crop Evol.* 60, 1763–1770. doi: 10.1007/s10722-013-9997-9

Geleta, M., Stymne, S., and Bryngelsson, T. (2011). Variation and inheritance of oil content and fatty acid composition in niger (*Guizotia abyssinica*). *J. food Compos. Anal.* 24, 995–1003. doi: 10.1016/j.jfca.2010.12.010

Glémin, S., Clément, Y., David, J., and Ressayre, A. (2014). GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. *Trends Genet.* 30, 263–270. doi: 10.1016/j.tig.2014.05.002

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.* 29:644. doi: 10.1038/nbt.1883

Grover, A., and Sharma, P. C. (2007). Microsatellite motifs with moderate GC content are clustered around genes on *Arabidopsis thaliana* chromosome 2. *In Silico Biol.* 7, 201–213.

Hildebrand, C. E., Torney, D. C., and Wagner, R. P. (1992). Informativeness of polymorphic DNA markers. *Los Alamos Sci.* 20, 100–102.

Hodgins, K. A., Lai, Z., Oliveira, L. O., Still, D. W., Scascitelli, M., Barker, M. S., et al. (2014). Genomics of compositae crops: reference transcriptome assemblies and evidence of hybridization with wild relatives. *Mol. Ecol. Resour.* 14, 166–177. doi: 10.1111/1755-0998.12163

Kalyandurg, P. B., Sundararajan, P., Dubey, M., Ghadamgahi, F., Zahid, M. A., Whisson, S., et al. (2021). Spray-induced gene silencing as a potential tool to control potato late blight disease. *Phytopathology* 111, 2168–2175. doi: 10.1094/PHYTO-02-21-0054-SC

Kashi, Y., King, D., and Soller, M. (1997). Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 13, 74–78. doi: 10.1016/S0168-9525(97)01008-1

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, 1179–1191. doi: 10.1111/1755-0998.12387

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

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

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

Li, Y.-C., Korol, A. B., Fahima, T., and Nevo, E. (2004). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.* 21, 991–1007. doi: 10.1093/molbev/msh073

Li, Y. L., and Liu, J. X. (2018). StructureSelector: a web—based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* 18, 176–177. doi: 10.1111/1755-0998.12719

Lu, J., Wang, S., Zhao, H., Liu, J., and Wang, H. (2012). Genetic linkage map of EST-SSR and SRAP markers in the endangered Chinese endemic herb *Dendrobium* (Orchidaceae). *Genet. Mol. Res.* 11, 4654–4667. doi: 10.4238/2012.December.21.1

Lu, J.-J., Kang, J.-Y., Feng, S.-G., Zhao, H.-Y., Liu, J.-J., and Wang, H.-Z. (2013). Transferability of SSR markers derived from *Dendrobium* nobile expressed sequence tags (ESTs) and their utilization in *Dendrobium* phylogeny analysis. *Sci. Hortic.* 158, 8–15. doi: 10.1016/j.scienta.2013.04.011

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

Mengistu, B., Gebrselassie, W., and Disasa, T. (2020). Diversity analysis in *Guizotia abyssinica* (L. f.) Cass. germplasms collected from Ethiopia. *Chem. Biomol. Eng.* 5, 8–14. doi: 10.11648/j.cbe.20200501.12

Misganaw, A., and Abera, S. (2017). Genetic diversity assessment of *Guzoita abyssinica* using EST derived simple sequence repeats (SSRs) markers. *Afr. J. Plant Sci.* 11, 79–85. doi: 10.5897/ajps2016.1512

Morgante, M., Hanafey, M., and Powell, W. (2002). Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30, 194–200. doi: 10.1038/ng822

Mun, J.-H., Kim, D.-J., Choi, H.-K., Gish, J., Debellé, F., Mudge, J., et al. (2006). Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. *Genetics* 172:2541–2555. doi: 10.1534/genetics.105.054791

Nemomissa, S., Bekele, E., and Dagne, K. (1999). Self-incompatibility system in the Ethiopian populations of *Guizotia abyssinica* (LF) Cass.(niger). *Sinet: Ethiop. J. Sci.* 22, 67–88.

Olmstead, J. W., Sebolt, A. M., Cabrera, A., Sooriyapathirana, S. S., Hammar, S., Iriarte, G., et al. (2008). Construction of an intra-specific sweet cherry (*Prunus avium* L.) genetic linkage map and synteny analysis with the Prunus reference map. *Tree Genet. Genomes* 4, 897–910. doi: 10.1007/s11295-008-0161-1

Peakall, R., and Smouse, P. E. (2006). GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Mol. Ecol. Notes* 6, 288–295. doi: 10.1093/bioinformatics/bts460

Petros, Y., Merker, A., and Zeleke, H. (2007). Analysis of genetic diversity of *Guizotia abyssinica* from Ethiopia using inter simple sequence repeat markers. *Hereditas* 144, 18–24. doi: 10.1111/j.2007.0018-0661.01969.x

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

Puechmaille, S. J. (2016). The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. *Mol. Ecol. Resour.* 16, 608–627. doi: 10.1111/1755-0998.12512

Qin, Z., Wang, Y., Wang, Q., Li, A., Hou, F., and Zhang, L. (2015). Evolution analysis of simple sequence repeats in plant genome. *PLoS One* 10:e0144108. doi: 10.1371/journal.pone.0144108

Raj, A., Stephens, M., and Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* 197, 573–589. doi: 10.1534/genetics.114.164350/-/DC1

Rannala, B., and Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. *Proc. Natl. Acad. Sci. U.S.A.* 94, 9197–9201. doi: 10.1073/pnas.94.17.9197

Serbessa Tolera, B., Dagne Woldegebriel, K., Teshome Gari, A., Geleta Dida, M., and Tesfaye Geletu, K. (2021). Analyses of genetic diversity and population structure of anchote (*Coccinia abyssinica* (Lam.) Cogn.) using newly developed EST-SSR markers. *Genet. Resour. Crop Evol.* 68, 2337–2350. doi: 10.1007/s10722-021-01132-5

Shete, S., Tiwari, H., and Elston, R. C. (2000). On estimating the heterozygosity and polymorphism information content value. *Theor. Popul. Biol.* 57, 265–271. doi: 10.1006/tpbi.2000.1452

Shiferaw, E., Pè, M., Porceddu, E., and Ponnaiah, M. (2012). Exploring the genetic diversity of Ethiopian grass pea (*Lathyrus sativus* L.) using EST-SSR markers. *Mol. Breed.* 30, 789–797. doi: 10.1007/s11032-011-9662-y

Shirasawa, K., Ishii, K., Kim, C., Ban, T., Suzuki, M., Ito, T., et al. (2013). Development of Capsicum EST–SSR markers for species identification and in silico mapping onto the tomato genome sequence. *Mol. Breed.* 31, 101–110. doi: 10.1007/s11032-012-9774-z

Singh, R., Ming, R., and Yu, Q. (2016). Comparative analysis of GC content variations in plant genomes. *Trop. Plant Biol.* 9, 136–149. doi: 10.1007/s12042-016-9165-4

Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., et al. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc. Natl. Acad. Sci. U.S.A.* 111, E4096–E4102. doi: 10.1073/pnas.1321152111

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., and McCouch, S. (2001). Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* 11, 1441–1452. doi: 10.1101/gr.184001

Teshome, A., Bryngelsson, T., Dagne, K., and Geleta, M. (2015). Assessment of genetic diversity in Ethiopian field pea (*Pisum sativum* L.) accessions with newly developed EST-SSR markers. *BMC Genet.* 16:102. doi: 10.1186/s1286301502615

Tóth, G., Gáspári, Z., and Jurka, J. (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10, 967–981. doi: 10.1101/gr.10.7.967

Tsehay, S., Ortiz, R., Johansson, E., Bekele, E., Tesfaye, K., Hammenhag, C., et al. (2020). New transcriptome-based snp markers for noug (*Guizotia abyssinica*) and their conversion to KASP markers for population genetics analyses. *Genes* 11:1373. doi: 10.3390/genes11111373

Van der Auwera, G. A., and O'Connor, B. D. (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. Sebastopol, CA: O'Reilly Media.

Wang, H.-C., Singer, G. A., and Hickey, D. A. (2004). Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* 21, 90–96. doi: 10.1093/molbev/msh003

Wang, J., Chen, C., Na, J.-K., Yu, Q., Hou, S., Paull, R. E., et al. (2008). Genome-wide comparative analyses of microsatellites in papaya. *Trop. Plant Biol.* 1, 278–292. doi: 10.1007/s12042-008-9024-z

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Zhao, Z., Guo, C., Sutharzan, S., Li, P., Echt, C. S., Zhang, J., et al. (2014). Genome-wide analysis of tandem repeats in plants and green algae. *G3 (Bethesda).* 4, 67–78. doi: 10.1534/g3.113.008524

Zhou, Q., Luo, D., Ma, L., Xie, W., Wang, Y., Wang, Y., et al. (2016). Development and cross-species transferability of EST-SSR markers in Siberian wildrye (*Elymus sibiricus* L.) using Illumina sequencing. *Sci. Rep.* 6:20549. doi: 10.1038/srep20549

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.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 01 August 2022
DOI 10.3389/fpls.2022.968618

OPEN ACCESS

EDITED BY
Andrés J. Cortés,
Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

REVIEWED BY
Yee-Shan Ku,
The Chinese University of Hong Kong, Hong Kong SAR, China
William M. Singer,
Virginia Tech, United States
Yingpeng Han,
Northeast Agricultural University, China
Daisuke Sekine,
National Agriculture and Food Research Organization (NARO), Japan

*CORRESPONDENCE
Yingshan Dong
email@uni.edu;
ysdong@cjaas.com

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

RECEIVED 14 June 2022
ACCEPTED 11 July 2022
PUBLISHED 01 August 2022

CITATION
Yuan B, Yuan C, Wang Y, Liu X, Qi G, Wang Y, Dong L, Zhao H, Li Y and Dong Y (2022) Identification of genetic loci conferring seed coat color based on a high-density map in soybean. *Front. Plant Sci.* 13:968618. doi: 10.3389/fpls.2022.968618

Identification of genetic loci conferring seed coat color based on a high-density map in soybean

Baoqi Yuan[1,2], Cuiping Yuan[2], Yumin Wang[2], Xiaodong Liu[3], Guangxun Qi[2], Yingnan Wang[2], Lingchao Dong[2], Hongkun Zhao[2], Yuqiu Li[2] and Yingshan Dong[1,2]*

[1]College of Agronomy, Jilin Agricultural University, Changchun, China, [2]Soybean Research Institute, Jilin Academy of Agricultural Sciences, National Engineering Research Center for Soybean, Changchun, China, [3]Crop Germplasm Institute, Jilin Academy of Agricultural Sciences, Changchun, China

Seed coat color is a typical evolutionary trait. Identification of the genetic loci that control seed coat color during the domestication of wild soybean could clarify the genetic variations between cultivated and wild soybean. We used 276 F_{10} recombinant inbred lines (RILs) from the cross between a cultivated soybean (JY47) and a wild soybean (ZYD00321) as the materials to identify the quantitative trait loci (QTLs) for seed coat color. We constructed a high-density genetic map using re-sequencing technology. The average distance between adjacent markers was 0.31 cM on this map, comprising 9,083 bin markers. We identified two stable QTLs (*qSC08* and *qSC11*) for seed coat color using this map, which, respectively, explained 21.933 and 26.934% of the phenotypic variation. Two candidate genes (*CHS3C* and *CHS4A*) in *qSC08* were identified according to the parental re-sequencing data and gene function annotations. Five genes (*LOC100786658*, *LOC100801691*, *LOC100806824*, *LOC100795475*, and *LOC100787559*) were predicted in the novel QTL *qSC11*, which, according to gene function annotations, might control seed coat color. This result could facilitate the identification of beneficial genes from wild soybean and provide useful information to clarify the genetic variations for seed coat color in cultivated and wild soybean.

KEYWORDS

soybean, re-sequencing, high-density genetic map, seed coat color, QTL

Introduction

Cultivated soybeans [*Glycine max* (L.) Merr.] were domesticated from wild soybeans by long-term targeted selection and improvement (Schmutz et al., 2010; Kim et al., 2012). The process of crop domestication encompasses a broad range of phenotypic changes throughout the multiple and continuous transition stages

(Meyer and Purugganan, 2013). To better clarify the genetic mechanisms of this process, many scientists have researched the whole-genome information of wild and cultivated soybean genomes to obtain a clearer picture of the modes of soybean domestication and diversification (Lam et al., 2010; Li et al., 2011; Li Y. H. et al., 2014; Zhou et al., 2015; Han et al., 2016; Sedivy et al., 2017; Liu et al., 2020). Individually, they assembled *de novo* different wild and cultivated soybean genomes and constructed a graph-based genome to reveal numerous genetic variations and gene fusion events. This novel information enables the search for candidate genes that have played important roles in soybean domestication and improvement.

Cultivated soybeans have a lower genetic diversity after domestication than their wild counterparts. The lower diversity has potentially resulted in the loss of genes that might be important in different environments (Hyten et al., 2006; Qi et al., 2014a). Therefore, wild soybeans that exhibit high allelic diversity may be an important resource for reintroduction into domesticated genes. The populations constructed by crossing cultivated and wild soybeans were more conducive to the identification of beneficial genes associated with the soybean domestication process.

Seed coat color is a typical domestication trait, evolving from black to yellow, green, brown and double color during soybean domestication from wild to cultivated (Han et al., 2016; Liu et al., 2021). Soybean seed coat color is mainly controlled by five genetic loci, designated as *I*, *R*, *T*, *W1*, and *O* classical genetic loci in previous reports (Senda et al., 2002a). The loci *I*, *R*, and *T* regulated seed coat color by controlling the synthesis of seed coat pigments (Song et al., 2016). In addition, Guiamet identified the cytoplasmic genetic locus *CytG* in plant chloroplasts (Guiamét et al., 2002). With the development of molecular biotechnology, more than 30 molecular marker loci on different chromosomes that control seed coat color in soybean have been detected. Researchers tended to construct the genetic map by mapping a population to provide an essential framework for the putative quantitative trait loci (QTLs) and genes (Githiri et al., 2007; Ohnishi et al., 2011; Qi et al., 2014b). Song et al. (2016) used a biparental population developed from the cross between two cultivated soybeans with yellow seed color and brown seed color to confirm the locus and in which different seed coat colors were further dissected into simple trait pairs. By genotyping the entire F_2 population using flanking markers located in fine-mapping regions, the genetic basis of seed coat color was dissected. Du et al. (2019) constructed a high-density linkage map of the recombinant inbred lines (RILs) population by using a specific length amplified fragment (SLAF) technique and determined the QTL of seed coat color and seed size for sesame. Zhang et al. (2019) used the RIL population derived from crossing 09A001 to identify the major and minor QTLs controlling seed coat color in *Brassica rapa* L. Li et al. (2021) identified the candidate genes regulating seed coat color in sesame using QTL mapping and transcriptome analysis by F_2 populations. Liu et al. (2021) used an improve wild soybean chromosome segment substitution line (CSSL) population from NN1138-2(*max*) × N24852(*soja*) to identify wild vs. cultivated gene alleles conferring seed coat color and days to flowering in soybean. They identified the same trait in different populations to identify consistent QTLs (Oyoo et al., 2011). A total of 20 genes were reported, and 15 of them were in the flavonoid metabolic pathway. The accumulation of flavonoid substances in dynamic equilibrium was the result of the interaction of transcription factors (Gillman et al., 2011; Cho et al., 2019; Jia et al., 2020). In addition, the interaction of some MYB (v-myb avian myeloblastosis viral oncogene homolog) transcription factors regulate the accumulation of flavonoid substances (Albert et al., 2014). Transcription factors such as GmMYB39 and GmMYB100 could negatively regulate the synthesis of isoflavones in soybean hairy roots (Liu et al., 2013; Yan et al., 2015). GmMYB58, GmMYB176, and GmMYB205 could positively regulate the synthesis of isoflavones (Yi et al., 2010; Han et al., 2017). The MYB transcription factors GmMYBA2 and GmMYBR are identified as transcriptional activators in a feedback loop to control the pigmentation of seed coat in soybeans (Gao et al., 2021). However, the genetic information controlling seed coat color during soybean domestication has not been completely elucidated and the transcriptional regulation relationship among the loci remains elusive.

To identify the loci and genes that controlling seed coat color, we used 276 F_{10} RIL populations developed from a cross between *Glycine max* and *Glycine soja* as the materials to construct a high-density genetic map by whole genome re-sequencing, map the additive QTLs, and predict candidate genes for seed coat color. The results of this study could facilitate the identification of beneficial genes from wild soybean and lead to a greater understanding of the process of soybean domestication.

Materials and methods

Plant materials and DNA extraction

The F_{10} RIL population (n = 276) was developed from a cross between Jiyu47 (JY47) and ZYD00321 using a single seed descent method. JY47 is an outstanding cultivated soybean with a yellow seed coat, ZYD00321 is a typical wild soybean with a black seed coat. The two parents and the RIL populations were planted in pots in the Gongzhuling Experiment Station at the Jilin Academy of Agricultural Sciences. We employed a planting pattern of two seeds per pot in three replicates to preserve the uniform density.

Fresh leaf tissue from the two parents and RIL individuals was collected at the flowering stage, immediately frozen in liquid nitrogen, then stored in a −80°C freezer. To obtain the high-quality DNA, the cetyltrimethylammonium bromide (CTAB) method was used to extract genomic DNA (Zhang et al., 2005).

The quality and concentration of the total genomic DNA were spectrophotometrically assessed by the optical density value (OD_{600} = 230/260, 260/280). The sequencing libraries were constructed following the manufacturer's instructions.

Genome re-sequencing and high-density genetic map construction

We performed whole-genome re-sequencing on RIL populations and the two parents to construct our high-density genetic map. Genome re-sequencing was constructed on the Illumina HiSeq2500 platform. We used an average sequencing depth of 20.00-fold in the two parents and 3.00-fold for individual RILs, and compared the sequence data with *Williams 82* (Glycine_max_v2.1) reference genome using the BWA package (Li and Durbin, 2009a) and combined the co-segregating markers which had been produced by the GATK process after comparison into bins using the HighMap software (Li et al., 2009b).

The HighMap software was used to analyze the linear arrangement of the bin markers within 20 linkage groups (LGs) and estimate the genetic distance between adjacent markers (Liu et al., 2014). The polymorphic single nucleotide polymorphisms (SNPs) were aligned with the reference genome and mapped onto 20 chromosomes (Chr). We calculated the MLOD scores between the polymorphic markers and filtered for MLOD values of less than 5. The HighMap software was used to calculate the map distances. SMOOTH (Van Ooijen, 2006) was applied to correct errors based on the parental contribution of the genotypes and a k-nearest neighbor algorithm was applied to impute missing genotypes. We mapped skewed markers by applying a multipoint method of maximum likelihood and estimated the map distances using the Kosambi mapping function in centimorgan (cm).

Phenotypic evaluation

We followed the "*Descriptors and Data Standard for soybean* (*Glycine* spp.)" (Qiu et al., 2006) to classify the traits and used the numbers 1–5 to represent the yellow, green, black, brown, and double color, respectively. The identified phenotypic data were collected and analyzed. We used Excel 2019 for statistics on all the phenotypic data and the software Graphpad prism 8.0 (Swift, 1997) for graphing.

Quantitative trait loci mapping and candidate genes prediction

The composite interval mapping (CIM) method of the *R/qtl* package (Arends et al., 2010) was used to detect additive QTLs for seed coat color. A total of 1,000 permutation tests at the 95% confidence level were used to set the logarithm-of-odds (LOD) threshold to detect significant QTLs (Churchill and Doerge, 1994). Based on 1,000 permutations, LOD = 5.356 was used to determine the presence of a putative QTL associated with the target trait in a particular genomic region. The QTLs were named as per the guidelines described (Swift, 1997). The sequences within the target QTLs were analyzed according to the *Williams 82* soybean reference genome sequence (Glycine_max_v2.1) in National Center for Biotechnology Information (NCBI). The physical positions of target intervals were aligned based on the same reference genome sequence. We obtained the SNPs and insertion-deletion (InDels) in the target intervals from the re-sequencing data and the genes with sequence variations between two parents to predict the candidate genes. We arranged the distributions of SNPs or InDels upstream, in the intragenic region and downstream.

We used the BLAST search on Soybase[1] to search for descriptions of the soybean genes. The CDS sequences from the QTL regions were retrieved from Phytozome[2]. The putative functions of the candidate genes were annotated based on the gene ontology (GO)[3] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[4] databases. We listed genes with similar functions or functional domains as the major candidate genes according to gene annotations and the functional analysis.

Results

Population sequencing and high-density genetic map construction

Recombinant inbred line populations derived from a cross between JY47 and ZYD00321 were sequenced on the Illumina HiSeq2500 platform to construct a high-density genetic map. A total of 20.85 Gb of clean data was obtained for JY47 and 20.99 Gb for ZYD00321 with 20.0-fold and 21.0-fold depth, respectively. The sequencing quality values (Q30) of the two parents were >93.00% and the GC content percentages (the proportion of Guanine and Cytosine of the whole genome) were, respectively, 35.76 and 35.84% (**Supplementary Table 1**).

A total of 2,612,708 SNPs between the parents were identified using the BWA package by comparing the sequencing data to the Williams 82 reference genome. The alignment efficiency was 96.26%. We obtained a total of 854.08 Gb of clean data with approximately 3.09-fold depth for each RIL. The average Q30 for the sequencing was 93.17% and the average GC content was 35.98% for each RIL. After filtering and

1 https://www.soybase.org
2 https://phytozome-next.jgi.doe.gov/
3 https://www.ebi.ac.uk/QuickGO/
4 http://www.kegg.jp/

quality assessment, 9083 bin markers without recombination events were used to construct the genetic map (Figure 1). The genotype of the RIL populations was generated to evaluate the genetic map quality. We used different colors to represent the origin of the different DNA fragments according to the physical location of 9,083 bin markers on 20 chromosomes (Supplementary Figure 1). It showed that this RIL population with a high recombination frequency was suitable for genetic analysis using marker-density linkage maps. A high-density genetic map with a total length of 2814.07 cM was constructed and the average distance between adjacent markers was 0.31 cM (Table 1). The genetic length of 20 LGs ranged from 103.69 cM (Chr11) to 160.19 cM (Chr10). The largest average distance was 1.01 cM on Chr17 with 135 bin markers and the smallest average density was 0.20 cM on Chr15 with 739 bin markers. The largest gap was mapped to Chr06 and was 18.82 cM in length. The proportion of gaps <5 cM between two markers was 94.33%.

To evaluate the collinearity between the genetic map and the soybean reference genome, 9083 bin markers were mapped to the soybean reference genome. A collinearity analysis showed that the order of markers on 20 chromosomes was consistent with the genome (Supplementary Figure 2). Consecutive curves between physical distances and genetic distances were observed except on Chr11 and Chr14. The Spearman coefficients of 20 LGs were >0.99 and collinearity was high at 99.80%, which indicated that the genetic and physical positions followed an identical order on this map. The high collinearity on our map indicated the genetic recombination rate was accurate and the gene annotation within QTL intervals was reliable.

Phenotypic variation of seed coat color

Seed coat color is a qualitative trait that is difficult for the environment to affect and had reached a state of purity for the F_{10} RIL population. Great phenotypic variation existed among the 276 RILs (Figure 2A). Five phenotypic types were found in the RIL population: black, brown, yellow, green, and double color (Figure 2B). The frequency distribution indicated that this RIL population was isolated for this trait and fulfilled the essential conditions for QTL localization. Of the 276 RILs, the brown seed coat was present in the largest quantity and the double color in the least quantity in the five classifications (Supplementary Table 2).

Quantitative trait loci mapping for seed coat color

Based on the constructed high-density genetic linkage map and the identified phenotypic analysis of seed coat color, we used the *R/qtl* package and the CIM program to identify QTLs associated with seed coat color in the RIL population (n = 276). The threshold of LOD scores for estimating the significant QTL effects was determined using 1,000 permutations. In total, two QTLs related to seed coat color designated as *qSC08* and *qSC11* were detected on Chr08 and Chr11, respectively (Table 2). The LOD score curves were constructed and sharp peaks spanning Chr08 and Chr11 were obtained (Figures 3A,B). The high phenotypic variance, respectively, explained by two QTLs ranged from 21.933 to 26.934% and the LOD score was 8.112 and 14.251. The additive effects of *qSC08* and *qSC11* were,

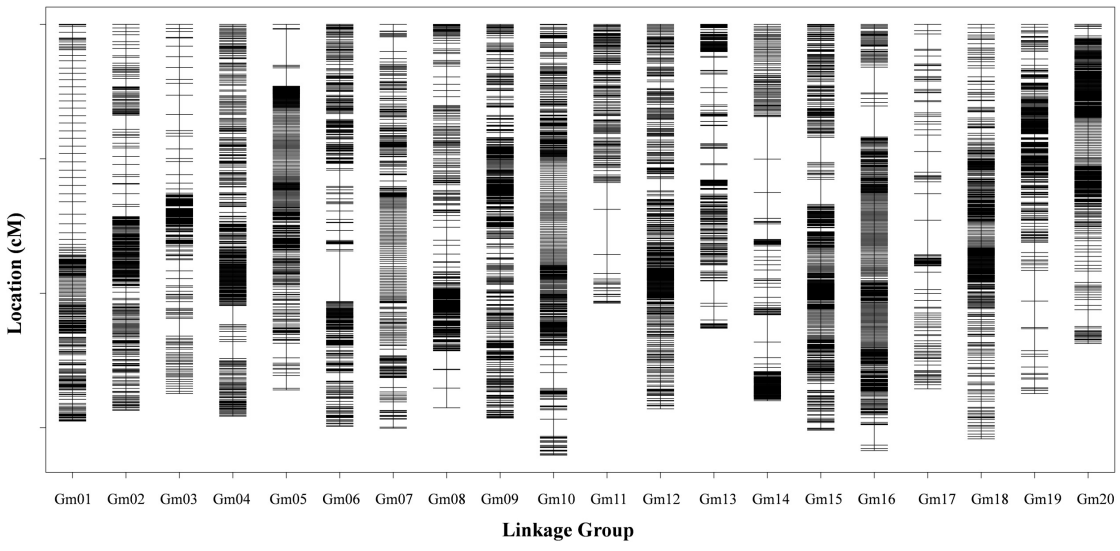


FIGURE 1
The soybean high-density genetic map. Bin markers are distributed on 20 chromosomes. The black bars in each linkage group represent the mapped bin markers. The linkage group number is shown on the *x*-axis and genetic distance is shown on the *y*-axis (cM is the unit).

TABLE 1 Characteristics of the high-density genetic map.

Linkage group ID	Total marker	Total distance (cM)	Average distance (cM)	Max gap (cM)	Gaps < 5 cM (%)
Chr01	389	147.54	0.38	4.63	94.33%
Chr02	472	143.55	0.30	5.22	99.58%
Chr03	264	137.34	0.52	7.20	98.86%
Chr04	590	145.77	0.25	7.07	99.66%
Chr05	514	135.98	0.27	13.47	99.42%
Chr06	448	149.49	0.33	18.82	98.43%
Chr07	345	150.19	0.44	5.44	99.71%
Chr08	372	142.60	0.38	7.31	99.19%
Chr09	645	146.43	0.23	3.08	100.00%
Chr10	665	160.19	0.24	6.36	99.40%
Chr11	219	103.69	0.48	16.78	98.62%
Chr12	604	143.01	0.24	4.56	100.00%
Chr13	267	113.04	0.42	8.30	96.24%
Chr14	335	140.00	0.42	15.76	98.20%
Chr15	739	150.91	0.20	9.34	99.73%
Chr16	499	158.52	0.32	11.53	99.40%
Chr17	135	135.57	1.01	12.86	95.52%
Chr18	596	154.16	0.26	3.26	100.00%
Chr19	448	137.35	0.31	11.38	99.33%
Chr20	537	118.73	0.22	6.24	98.88%
Total	9083	2814.07	0.31	18.82	94.33%

respectively, −0.616 and −0.683 and the beneficial alleles of two major and stable QTLs were derived mainly from the male parent ZYD00321 (Table 2). The results indicated that two loci *qSC08* and *qSC11* had a powerful effect on the seed coat color.

Gene annotation and candidate genes prediction

To validate the QTL mapping results, we annotated and analyzed the potential genes within the QTL intervals by comparing the genome interval regions within the QTLs with the reference genome sequences. The 0.326 cM physical interval for *qSC08* represents approximately 140 kb in the reference genome and contains 10 candidate genes according to the annotation of *Williams 82* (Figure 3C). We analyzed the SNPs and InDels based on the whole genome re-sequencing data of both parents (JY47 and ZYD00321) to understand the genetic variations of these genes. In *qSC08*, 8 of 10 genes possessed SNPs or InDels. In total, 254 SNPs and 51 InDels were detected among 8 genes (Supplementary Table 3). Among these variations, a percentage of 44.59% (136/305) were located outside of the genes, including the scope within or beyond 5 kb upstream and downstream of the transcription start and stop sites. A percentage of 44.59% variations were located in the intergenic region. Non-synonymous variations with a percentage of 9.83% (30/305) were found in the coding sequence among the intragenic region (Figure 4A and Supplementary Table 3). Additionally, we annotated the functions of 8 variant genes based on the GO and KEGG databases to anchor the candidate genes for seed coat color in soybean (Supplementary Table 4). The results indicated that *LOC100789075* (*GLYMA_08G110300)* and *LOC100779649* (*GLYMA_08G110700)* might be involved in the response to seed coat color in soybean. They encoded chalcone synthase (*CHS3C* and *CHS4A*) and were involved in the flavonoid biosynthetic pathway. Chalcone synthase (*CHS*) is a key enzyme in the branch of the phenylpropanoid pathway leading to the biosynthesis of flavonoid pigments including anthocyanins. A sequence comparison analysis between the parents supported the above prediction. SNP or InDel variations between both parents were found in the upstream regions of the two genes; one InDel and three SNPs for *GLYMA_08G110300* and eight SNPs for *GLYMA_08G110700* (Figure 3D). Based on the functional annotation of candidate genes and sequence alignment analysis between the two parents, we predicted *GLYMA_08G110300* and *GLYMA_08G110700* as candidate genes that controlled seed coat color in soybean.

Based on the Williams 82 soybean reference genome, a total of 281 genes occupied the novel *qSC11* were identified. We analyzed the SNPs/InDels based on the whole genome re-sequencing data of two parents to understand the genetic variations of these genes. A total of 256 variant genes contained 9,996 SNPs and 2,191 InDels in *qSC11* were identified as the candidate genes for seed coat color (Supplementary Table 5).

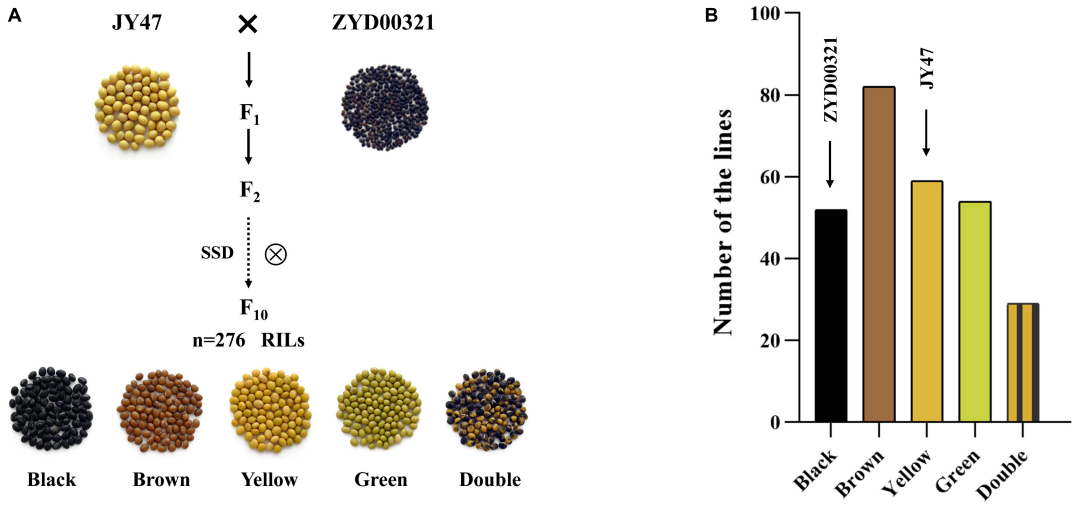


FIGURE 2
Phenotypic identification of parents and RIL populations. **(A)** The characteristics of the seed coat color of JY47 and ZYD00321 and RIL populations. **(B)** Frequency distribution of seed coat color for the 276 RILs.

TABLE 2 Two QTLs for seed coat color in RIL populations.

Name	Chr	Genetic interval	Physical interval	Marker interval	LOD	ADD	PVE (%)
qSC08	D1b	43.225–43.551	8449385–8588340	Block89587–Block89563	14.251	−0.616	21.933
qSC11	F	102.961–103.143	11236206–22112949	Block125748–Block126245	8.112	−0.683	26.934

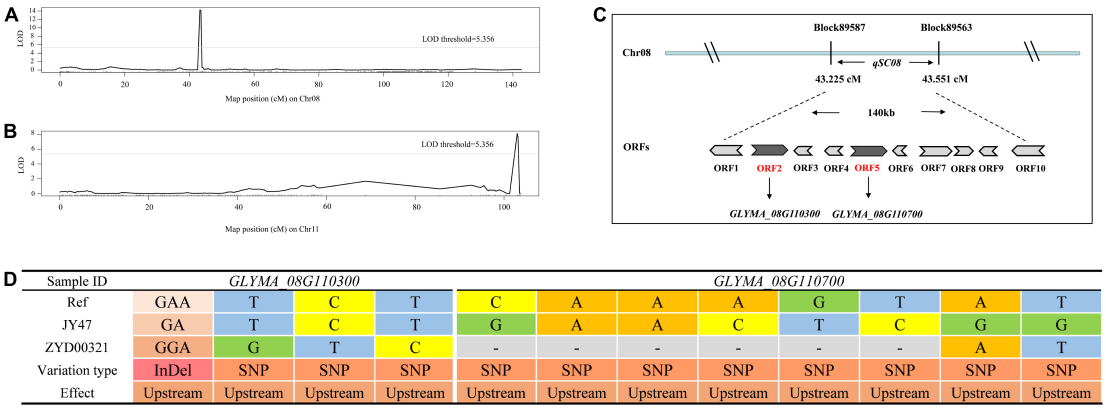
Among these variations, a substantial portion (68.48%) was located outside of the genes, including the scope within or beyond 5 kb upstream and downstream of the transcription start and stop sited (Figure 4B). Only 29.58% of variations were located in the intragenic region (Figure 4A and Supplementary Table 5). Non-synonymous variations with a percentage of 15.27% (467/3059) were found in the coding sequence among the intragenic region (Figure 4B and Supplementary Table 5). It was speculated that the key genes regulating seed coat color existed in the target intervals from the mass of genetic variations between the parents.

The gene functions of 256 variant genes were annotated to anchor the candidate genes for soybean seed coat color based on GO and KEGG databases, among which, only 131 genes were annotated (Supplementary Table 6). A total of 122 genes were annotated in the GO database as cellular components, molecular functions and biological processes (Supplementary Figure 3 and Supplementary Table 6), and 71 genes were detected in the KEGG database (Supplementary Table 6). According to the gene annotation results, five genes were annotated that might involve the biosynthetic pathway that controlled seed coat color, including *LOC100786658*, *LOC100801691*, *LOC100806824*, *LOC100795475*, and *LOC100787559* (Figure 5A and Supplementary Table 6). Of these, *LOC100786658* and *LOC100801691* encoded xanthoxin dehydrogenase and are involved in carotenoid biosynthesis. *LOC100806824* and *LOC100795475* encoded photosystem I reaction center subunit VI and protein TIC110 from chloroplast, respectively. *LOC100787559* encodes cytochrome P450. All the five genes existed with SNP or InDel variations in the coding region; *LOC100786658*, *LOC100801691*, and *LOC100787559* existed with SNP or InDel variations in the upstream, coding region and downstream between two parents; *LOC100806824* only existed two SNPs in the intragenic region; *LOC100795475* existed six SNPs and one InDels in the intragenic region, severally (Figure 5B and Supplementary Table 7). Among which, the candidate genes *LOC100795475* had two non-synonymous coding variations (Act/Gct, cGt/cAt) between two parents (Supplementary Table 7). All the five candidate genes were annotated as affecting the composition and content of pigments from seed coats in different ways.

Discussion

The high-density genetic map for quantitative trait loci mapping

A proper marker density for high-density genetic maps could provide an essential framework for QTL fine mapping (Gutierrez-Gonzalez et al., 2011; Qi et al., 2014a). In previous



D

Sample ID	GLYMA_08G110300				GLYMA_08G110700							
Ref	GAA	T	C	T	C	A	A	A	G	T	A	T
JY47	GA	T	C	T	G	A	A	C	T	C	G	G
ZYD00321	GGA	G	T	C	-	-	-	-	-	-	A	T
Variation type	InDel	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP	SNP
Effect	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream	Upstream

FIGURE 3
Fine mapping of two QTLs. **(A)** Mapping of QTLs for seed coat color on Chr08. The curves indicate the physical position of markers against the LOD score of the QTLs detected on Chr08. **(B)** Mapping of the QTLs for seed coat color on Chr11. The curves indicate the physical position of the markers against the LOD score of the QTLs detected on Chr11. **(C)** Distribution of tightly linked markers and plausible candidate genes for seed coat color on *qSC08* in soybean. **(D)** SNP/InDels variants based on re-sequencing data from both parents of *GLYMA_08G110300* and *GLYMA_08G110700*.

studies, the genetic maps constructed with restriction fragment length polymorphism (RFLP) and SSR markers have drawbacks of relatively few markers and large gaps, which limited the efficiency and accuracy of QTL mapping. With the completion of the whole genome sequencing of Williams 82 (the reference genome in this study) and the rapid development of sequencing technology, SNP markers have become widely used to construct genetic maps in plants (Hyten et al., 2008). Cai et al. (2018) used a high-density genetic linkage map containing 3,469 recombination bin markers based on 0.2 × RAD-seq technology to map QTLs for isoflavone content. Han et al. (2019) constructed a high-density genetic map using 260 RILs derived from the cultivars of Heihe43 and Heihe18, and the constructed map contained 4,953 SLAF markers spanning 1478.86 cM with an average distance between adjacent markers of 0.53 cM. Chu et al. (2021) reported on a genetic linkage map constructed by polymorphic 2,234 SNP markers from a SoySNP6K array, which covered a total of 4229.01 cM genetic distance with an average distance of 1.89 cM. Tian et al. (2022) constructed a high-density genetic map by re-sequencing technology, which contained a total of 4,011 recombination bin markers with an average distance of 0.78 cM in the entire RILs population. In this study, we used F_{10} RILs from the cross between JY47 and ZYD00321 to construct the high-density genetic map that contained 9,083 bin markers with an average distance of 0.31 cM between adjacent markers.

Although the resolution of genetic maps has been improved by increasing marker density, it has been limited by a Linkage disequilibrium (LD) in soybean that is significantly higher than in other plants (Lam et al., 2010; Gutierrez-Gonzalez et al., 2011). Because the average LD of cultivated soybean is approximately 150 kb, at least 6,300 distributed markers could theoretically fulfill a high-density genetic map (Li B. et al., 2014). In this study, we used the re-sequencing technology with high efficiency and capacity to construct a high-density genetic map. The number of bin markers was significantly higher than the theoretical value of the genetic map. Compared with the previously constructed high-density genetic maps, our map presented the characteristics of more markers (9083 bin markers), a smaller average genetic distance (0.31 cM) and higher collinearity (99.80%), which effectively eliminated the drawback of a large gap. These results indicated that the drawback of the high link disequilibrium could be avoided and fulfill the high-density genetic map could be achieved. Moreover, the use of RIL population with a wide range of variation can enhance our understanding of molecular mechanism evolution and genetic regulation, and also help to identify more QTLs regulating seed coat color in soybeans.

Identification and evaluation of quantitative trait locis for seed coat color

High-generation RIL populations were excellent materials for QTL localization. In this study, we identified two major QTLs for seed coat color, *qSC08* and *qSC11*, by the F_{10} RIL population with significant isolation for seed coat color. However, the *qSC11* presented a much broader interval compared with *qSC08*, we speculate the reason was the non-uniform distribution of the 9,083 bin markers on 20 chromosomes. In comparison to the previous results, *qSC08* was mapped into a much smaller region (Ohnishi et al., 2011; Qi et al., 2014a; Liu et al., 2021). Senda et al. (2002b) and Song et al. (2016) had shown five classical genetic loci *I*, *R*, *T*, *W1*, and *O*. Coincidently, *qSC08* was located precisely within the classical

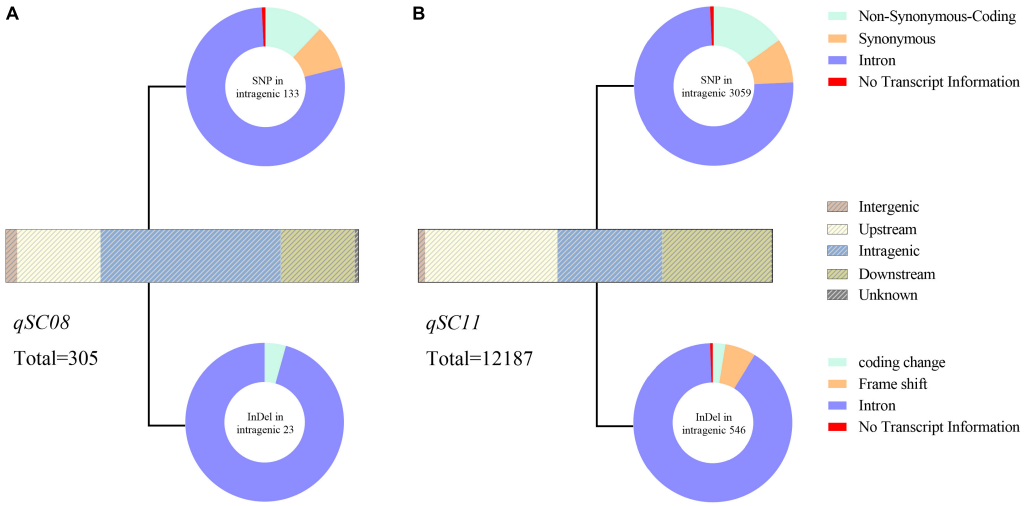


FIGURE 4
Analysis of SNPs and InDels between two parents in *qSC08* **(A)** and *qSC11* **(B)**. Strip-shape charts show the distribution of SNPs and InDels in different genomic regions. The upstream and downstream represent the 5 kb within the region of transcription start and stop sites, respectively. Pie charts show the effects of SNP (upside) and InDel (underside) in the intragenic regions. And the corresponding quantity of SNP or InDel is labeled near the pie chart.

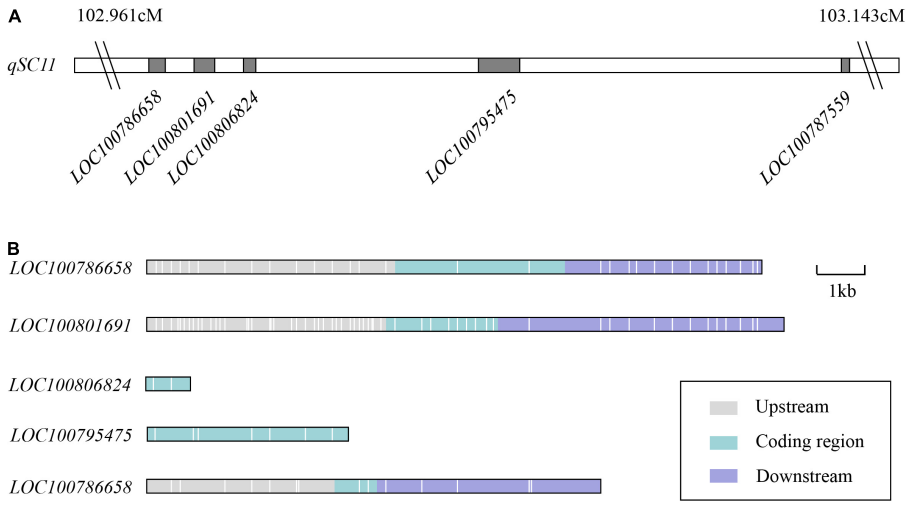


FIGURE 5
Distribution of the five candidate genes for seed coat color on *qSC11*. **(A)** Distribution of the candidate genes on *qSC11*. **(B)** Variation distribution of the candidate genes on *qSC11*. The upstream and downstream represent the 5 kb within the region of transcription start and stop sites, respectively. The regions are markers of different colors. Each white line represents one SNP/InDel variant.

genetic *I* locus that controlled seed coat color by regulating the distribution of anthocyanin and proanthocyanidin from seed coat (Todd and Vodkin, 1996). The *I* locus was located in the chalcone synthase *CHS* gene-rich region (Clough et al., 2004). The chalcone synthase gene family of soybean includes *CHS1*, *CHS2*, *CHS3*, *CHS4*, *CHS5*, *CHS6*, *CHS7*, *CHS8*, and *CHS9* (Clough et al., 2004). Notably, the candidate gene *CHS4* (*GLYMA_08G110700*) in our study is a member of the chalcone synthase gene family (Senda et al., 2002b). The clear conclusion was that the *qSC08* locus had a powerful effect on seed coat color in soybean and also demonstrated the accuracy and reliability of this study.

Three QTLs for seed coat color were detected on Chr11 in the previous studies. Of these, the classical genetic locus *K1* controlled the distribution of pigment in the saddle region, regulating seed coat color in soybean (Cho et al., 2017). The *D2* locus determines a yellow or green coat according to the chlorophyll content in the seed

coat (Fang et al., 2014). Kovinich et al. (2011, 2012) detected a locus within the physical interval 1992156–1993544 on Chr11 and identified the candidate gene *Glyma.11g027700* that encoded anthocyanidin synthase *ANS3*. The previous reports had no QTL for seed coat color in the physical interval 11236206–22112949 on Chr11, so the target region *qSC11* was a novel QTL. We predicted five candidate genes from *qSC11*. Of these, *LOC100801691* and *LOC100786658* encoding xanthoxin dehydrogenase and are involved in carotenoid biosynthesis. Carotenoids are the second most abundant natural pigments with more than 750 members. The color of carotenoids varies from colorless to yellow, orange, and red with variations reflected in plants (Nisar et al., 2015). *LOC100806824* and *LOC100795475* encoded photosystem I reaction center subunit VI and protein TIC110 from the chloroplast. During photomorphogenesis, the chlorophyll and carotenoid compounds are promoted in a coordinated manner in the development of photosynthesis (Welsch et al., 2000; Rodríguez-Villalón et al., 2009). In chloroplasts, most carotenoid biosynthetic genes are activated during light-triggered de-etiolation (Giuliano et al., 2008; Rodríguez-Concepción, 2010), and it indirectly affecting the accumulation of pigments from the individual tissues of the plants. *LOC100787559* encodes cytochrome P450, which plays an important role in flavonoid biosynthesis and the principal cytochromes in plants (Tanaka, 2006; Severin et al., 2010; Waese et al., 2017). In the present study, SNP and InDel variations were also observed between the two parents in the genomic sequences of the five candidate genes, including the regions 5-kb downstream and upstream of the genes and the coding regions (**Figure 5** and **Supplementary Table 7**). It was also found that the candidate genes *LOC100795475* occurred with non-synonymous coding variations (Act/Gct, cGt/cAt) between the two parents. Therefore, it was speculated that these genes might be the key genes responsible for soybean seed coat color. The cloning and functional analyses of these candidate genes will be conducted in the future, which will help to expound the genetic variations between wild and cultivated soybean more thoroughly regarding seed coat color.

Identification of important loci and genes in wild soybean

Cultivated soybeans were domesticated from wild soybeans *via* long-term selection and improvement (Sedivy et al., 2017). In previous studies, researchers usually located and analyzed target traits by constructing populations. Cho et al. (2021) used a population derived from a cross between a Korean cultivar and IT162669 to identify QTLs conferring salt tolerance in soybean. Githiri et al. (2007) used an F_2 population derived from a cross between two cultivated soybeans to identify five QTLs for pigmentation. Ohnishi et al. (2011) used two sets of RILs between two cultivars to identify minor QTLs for seed coat color. However, such procedures could not fully reflect the changes in seed coat color and might miss some vital genetic information during the domestication process.

Identification of genes and alleles from wild germplasm associated with seed coat color could allow deeper insight into the process of the changes in this trait during soybean domestication (Hyten et al., 2006; Lam et al., 2010; Zhuang et al., 2022). ZYD00321 is a typical wild soybean with a black seed coat, a small seed and a vining growth habit. We used the RIL population derived from JY47 (*Glycine max*) and ZYD00321 (*Glycine soja*) to identify QTLs and genes for seed coat color. Identification of the source of the beneficial alleles on each QTL is the prerequisite for the QTLs application to molecular breeding and crop improvement (Wang et al., 2007). In this study, *qSC08* and the novel interval *qSC11* showed consistent ADD (−0.616 and −0.683) and similar PVE (21.933 and 26.934%), which indicated that both beneficial alleles were derived from the wild soybean ZYD00321 and demonstrated that a wild soybean with a black seed coat had a crucial role in producing different seed coat colors, facilitating the identification of superior genes in the domestication process (Alkan et al., 2011; Li et al., 2011). The intact and accurate genomic information we obtained for wild germplasm was beneficial for identifying QTLs and conducting association studies on seed coat color (Kim et al., 2010; Xie et al., 2019).

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/, PRJNA848661.

Author contributions

YD and BY conceived and designed the experiments and methods. BY performed the experiments and drafted the manuscript. YD and CY reviewed the manuscript. YuW provided the material. GQ, YiW, and LD performed the field management. BY, XL, HZ, and YL analyzed the data and revised the manuscript. All authors read and approved the manuscript.

Funding

This research was supported by the Agricultural Science and Technology Innovation Project of Jilin Province (CXGC2017JQ018 and CXGC2018ZY010), the Jilin Academy of Agricultural Sciences Balance Fund Project (y81980401), and the National Key R&D Program of China (2021YFD1200103-1).

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

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.2022.968618/full#supplementary-material

References

Albert, N. W., Davies, K. M., and Schwinn, K. E. (2014). Gene regulation networks generate diverse pigmentation patterns in plants. *Plant Signal Behav.* 9, 962–980. doi: 10.4161/psb.29526

Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958

Arends, D., Prins, P., Jansen, R. C., and Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics* 26, 2990–2992. doi: 10.1093/bioinformatics/btq565

Cai, Z., Cheng, Y., Ma, Z., Liu, X., Ma, Q., Xia, Q., et al. (2018). Fine-mapping of QTLs for individual and total isoflavone content in soybean (*Glycine max L.*) using a high-density genetic map. *Theor. Appl. Genet.* 131, 555–568. doi: 10.1007/s00122-017-3018-x

Cho, K. H., Kim, M. Y., Kwon, H., Yang, X., and Lee, S. H. (2021). Novel QTL identification and candidate gene analysis for enhancing salt tolerance in soybean (*Glycine max* (L.) Merr.). *Plant Sci.* 313:111085. doi: 10.1016/j.plantsci.2021.111085

Cho, Y. B., Jones, S. I., and Vodkin, L. O. (2017). Mutations in Argonaute5 illuminate epistatic interactions of the *K1* and *I* loci leading to saddle seed color patterns in *Glycine max*. *Plant Cell* 29, 708–725. doi: 10.1105/tpc.17.00162

Cho, Y. B., Jones, S. I., and Vodkin, L. O. (2019). Nonallelic homologous recombination events responsible for copy number variation within an RNA silencing locus. *Plant Direct* 3:e00162. doi: 10.1002/pld3.162

Chu, J., Li, W., Piao, D., Lin, F., Huo, X., Zhang, H., et al. (2021). Identification of a major QTL related to resistance to soybean mosaic virus in diverse soybean genetic populations. *Euphytica* 217, 1–11. doi: 10.1007/s10681-021-02907-8

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

Clough, S. J., Tuteja, J. H., Li, M., Marek, L. F., Shoemaker, R. C., and Vodkin, L. O. (2004). Features of a 103-kb gene-rich region in soybean include an inverted perfect repeat cluster of *CHS* genes comprising the *I* locus. *Genome* 47, 819–831. doi: 10.1139/g04-049

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

Fang, C., Li, C., Li, W., Wang, Z., and Tian, Z. (2014). Concerted evolution of *d1* and *d2* to regulate chlorophyll degradation in soybean. *Plant J.* 77, 700–712. doi: 10.1111/tpj.12419

Gao, R., Han, T., Xun, H., Zeng, X., Li, P., Li, Y., et al. (2021). MYB transcription factors GmMYBA2 and GmMYBR function in a feedback loop to control pigmentation of seed coat in soybean. *J. Exp. Bot.* 72, 4401–4418. doi: 10.1093/jxb/erab152

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

Githiri, S. M., Yang, D., Khan, N. A., Xu, D., Komatsuda, T., and Takahashi, R. (2007). QTL analysis of low temperature–induced browning in soybean seed coats. *J. Heredi.* 98, 360–366. doi: 10.1093/jhered/esm042

Giuliano, G., Tavazza, R., Diretto, G., Beyer, P., and Taylor, M. A. (2008). Metabolic engineering of carotenoid biosynthesis in plants. *Trends Biotechnol.* 26, 139–145. doi: 10.1016/j.tibtech.2007.12.003

Guiamét, J. J., Tyystjärvi, E., Tyystjärvi, T., John, I., Kairavuo, M., Pichersky, E., et al. (2002). Photoinhibition and loss of photosystem II reaction centre proteins during senescence of soybean leaves. Enhancement of photoinhibition by the 'stay-green'mutation *cytG*. *Physiol. Plantarum* 115, 468–478. doi: 10.1034/j.1399-3054.2002.1150317.x

Gutierrez-Gonzalez, J. J., Vuong, T. D., Zhong, R., Yu, O., Lee, J. D., Shannon, G., et al. (2011). Major locus and other novel additive and epistatic loci involved in modulation of isoflavone concentration in soybean seeds. *Theor. Appl. Genet.* 123, 1375–1385. doi: 10.1007/s00122-011-1673-x

Han, J., Han, D., Guo, Y., Yan, H., Wei, Z., Tian, Y., et al. (2019). QTL mapping pod dehiscence resistance in soybean (*Glycine max* L. Merr.) using specific-locus amplified fragment sequencing. *Theor. Appl. Genet.* 132, 2253–2272. doi: 10.1007/s00122-019-03352-x

Han, X., Yin, Q., Liu, J., Jiang, W., Di, S., and Pang, Y. (2017). GmMYB58 and GmMYB205 are seed-specific activators for isoflavonoid biosynthesis in *Glycine max*. *Plant Cell Rep.* 36, 1889–1902. doi: 10.1007/s00299-017-2203-3

Han, Y., Zhao, X., Liu, D., Li, Y., Lightfoot, D. A., Yang, Z., et al. (2016). Domestication footprints anchor genomic regions of agronomic importance in soybeans. *New Phytol.* 209, 871–884. doi: 10.1111/nph.13626

Hyten, D. L., Song, Q., Choi, I. Y., Yoon, M. S., Specht, J. E., Matukumalli, L. K., et al. (2008). High-throughput genotyping with the GoldenGate assay in the complex genome of soybean. *Theor. Appl. Genet.* 116, 945–952. doi: 10.1007/s00122-008-0726-2

Hyten, D. L., Song, Q., Zhu, Y., Choi, I. Y., Nelson, R. L., Costa, J. M., et al. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Nat. Acad. Sci. U.S.A.* 103, 16666–16671. doi: 10.1073/pnas.0604379103

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

Kim, M. Y., Lee, S., Van, K., Kim, T. H., Jeong, S. C., Choi, I. Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc. Nat. Acad. Sci. U.S.A.* 107, 22032–22037. doi: 10.1073/pnas.1009526107

Kim, M. Y., Van, K., Kang, Y. J., Kim, K. H., and Lee, S. H. (2012). Tracing soybean domestication history: From nucleotide to genome. *Breeding Sci.* 61, 445–452. doi: 10.1270/jsbbs.61.445

Kovinich, N., Saleem, A., Arnason, J. T., and Miki, B. (2011). Combined analysis of transcriptome and metabolite data reveals extensive differences between black and brown nearly-isogenic soybean (*Glycine max*) seed coats enabling the identification of pigment isogenes. *BMC Genomics* 12:381. doi: 10.1186/1471-2164-12-381

Kovinich, N., Saleem, A., Rintoul, T. L., Brown, D. C., Arnason, J. T., and Miki, B. (2012). Coloring genetically modified soybean grains with anthocyanins by suppression of the proanthocyanidin genes *ANR1* and *ANR2*. *Transgenic Res.* 21, 757–771. doi: 10.1007/s11248-011-9566-y

Lam, H. M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F. L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42, 1053–1059. doi: 10.1038/ng.715

Li, B., Tian, L., Zhang, J., Huang, L., Han, F., Yan, S., et al. (2014). Construction of a high-density genetic map based on large-scale markers developed by specific length amplified fragment sequencing (SLAF-seq) and its application to QTL analysis for isoflavone content in *Glycine max*. *BMC Genomics* 15:1086. doi: 10.1186/1471-2164-15-1086

Li, C., Duan, Y., Miao, H., Ju, M., Wei, L., and Zhang, H. (2021). Identification of candidate genes regulating the seed coat color trait in sesame (*Sesamum indicum* L.) using an integrated approach of QTL mapping and transcriptome analysis. *Front. Genet.* 12:700469. doi: 10.3389/fgene.2021.700469

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

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

Li, Y., Zheng, H., Luo, R., Wu, H., Zhu, H., Li, R., et al. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.* 29, 723–730. doi: 10.1038/nbt.1904

Li, Y. H., Zhou, G., Ma, J., Jiang, W., Jin, L. G., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052. doi: 10.1038/nbt.2979

Liu, C., Chen, X., Wang, W., Hu, X., Han, W., He, Q., et al. (2021). Identifying wild versus cultivated gene-alleles conferring seed coat color and days to flowering in soybean. *Int. J. mol. Sci.* 22:1559. doi: 10.3390/ijms22041559

Liu, D., Ma, C., Hong, W., Huang, L., Liu, M., Liu, H., et al. (2014). Construction and analysis of high-density linkage map using high-throughput sequencing data. *PLoS One* 9:e98855. doi: 10.1371/journal.pone.0098855

Liu, X., Yuan, L., Xu, L., Xu, Z., Huang, Y., He, X., et al. (2013). Over-expression of *GmMYB39* leads to an inhibition of the isoflavonoid biosynthesis in soybean (*Glycine max, L). Plant Biotechnol. Rep.* 7, 445–455. doi: 10.1007/s11816-013-0283-2

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182, 162–176. doi: 10.1016/j.cell.2020.05.023

Meyer, R. S., and Purugganan, M. D. (2013). Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* 14, 840–852. doi: 10.1038/nrg3605

Nisar, N., Li, L., Lu, S., Khin, N. C., and Pogson, B. J. (2015). Carotenoid metabolism in plants. *Mol. Plant* 8, 68–82. doi: 10.1016/j.molp.2014.12.007

Ohnishi, S., Funatsuki, H., Kasai, A., Kurauchi, T., Yamaguchi, N., Takeuchi, T., et al. (2011). Variation of *GmIRCHS* (*Glycine max* inverted-repeat *CHS* pseudogene) is related to tolerance of low temperature-induced seed coat discoloration in yellow soybean. *Theor. Appl. Genet.* 122, 633–642. doi: 10.1007/s00122-010-1475-6

Oyoo, M. E., Benitez, E. R., Kurosaki, H., Ohnishi, S., Miyoshi, T., Kiribuchi-Otobe, C., et al. (2011). QTL analysis of soybean seed coat discoloration associated with *II TT* genotype. *Crop Sci.* 51, 464–469. doi: 10.2135/cropsci2010.02.0121

Qi, X., Li, M. W., Xie, M., Liu, X., Ni, M., Shao, G., et al. (2014a). Identification of a novel salt tolerance gene in wild soybean by whole-genome sequencing. *Nat. Commun.* 5:5340. doi: 10.1038/ncomms5340

Qi, Z., Huang, L., Zhu, R., Xin, D., Liu, C., Han, X., et al. (2014b). A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PLoS One* 9:e104871. doi: 10.1371/journal.pone.0104871

Qiu, L., Chang, R., et al. (2006). *Descriptors and Data Stan-dard for Soybean (Glycine spp)*. Beijing: Chinese Agriculture Press.

Rodríguez-Concepción, M. (2010). Supply of precursors for carotenoid biosynthesis in plants. *Arch. Biochem. Biophys.* 504, 118–122. doi: 10.1016/j.abb.2010.06.016

Rodríguez-Villalón, A., Gas, E., and Rodríguez-Concepción, M. (2009). Colors in the dark: a model for the regulation of carotenoid biosynthesis in etioplasts. *Plant Signal Behav.* 4, 965–967. doi: 10.4161/psb.4.10.9672

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

Sedivy, E. J., Wu, F., and Hanzawa, Y. (2017). Soybean domestication: the origin, genetic architecture and molecular bases. *New Phytol.* 214, 539–553. doi: 10.1111/nph.14418

Senda, M., Kasai, A., Yumoto, S., Akada, S., Ishikawa, R., Harada, T., et al. (2002a). Sequence divergence at chalcone synthase gene in pigmented seed coat soybean mutants of the Inhibitor locus. *Genes Genet. Syst.* 77, 341–350. doi: 10.1266/ggs.77.341

Senda, M., Jumonji, A., Yumoto, S., Ishikawa, R., Harada, T., Niizeki, M., et al. (2002b). Analysis of the duplicated *CHS1* gene related to the suppression of the seed coat pigmentation in yellow soybeans. *Theor. Appl. Genet.* 104, 1086–1091. doi: 10.1007/s00122-001-0801-4

Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). Rna-seq atlas of glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* 10:160. doi: 10.1186/1471-2229-10-160

Song, J., Liu, Z., Hong, H., Ma, Y., Tian, L., Li, X., et al. (2016). Identification and validation of loci governing seed coat color by combining association mapping and bulk segregation analysis in soybean. *PLoS One* 11:e0159064. doi: 10.1371/journal.pone.0159064

Swift, M. L. (1997). GraphPad prism, data analysis, and scientific graphing. *J. Chem. Inf. Comput. Sci.* 37, 411–412. doi: 10.1021/ci960402j

Tanaka, Y. (2006). Flower colour and cytochromes p450. *Phytochem. Rev.* 5, 283–291. doi: 10.1007/s11101-006-9003-7

Tian, Y., Yang, L., Lu, H. F., Zhang, B., Li, Y. F., Liu, C., et al. (2022). QTL analysis for plant height and fine mapping of two environmentally stable QTLs with major effects in soybean. *J. Integr. Agr.* 21, 933–946. doi: 10.1016/S2095-3119(21)63693-6

Todd, J. J., and Vodkin, L. O. (1996). Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. *Plant Cell* 8, 687–699. doi: 10.1105/tpc.8.4.687

Van Ooijen, J. (2006). *Software for the calculation of genetic linkage maps in experimental populations Kyazma BV[J]*. Wageningen: Kyazma BV.

Waese, J., Fan, J., Pasha, A., Yu, H., Fucile, G., Shi, R., et al. (2017). Eplant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *Plant Cell.* 29, 1806–1821. doi: 10.1105/tpc.17.00073

Wang, J., Wan, X., Li, H., Pfeiffer, W. H., Crouch, J., and Wan, J. (2007). Application of identified QTL-marker associations in rice quality improvement through a design-breeding approach. *Theor. Appl. Genet.* 115, 87–100. doi: 10.1007/s00122-007-0545-x

Welsch, R., Beyer, P., Hugueney, P., Kleinig, H., and von Lintig, J. (2000). Regulation and activation of phytoene synthase, a key enzyme in carotenoid biosynthesis, during photomorphogenesis. *Planta* 211, 846–854. doi: 10.1007/s004250000352

Xie, M., Chung, C. Y. L., Li, M. W., Wong, F. L., Wang, X., Liu, A. L., et al. (2019). A reference-grade wild soybean genome. *Nat. Commun.* 10:1216. doi: 10.1038/s41467-019-09142-9

Yan, J., Wang, B., Zhong, Y., Yao, L., Cheng, L., and Wu, T. (2015). The soybean R2R3 MYB transcription factor *GmMYB100* negatively regulates plant flavonoid biosynthesis. *Plant Mol. Biol.* 89, 35–48. doi: 10.1007/s11103-015-0349-3

Yi, J., Derynck, M. R., Li, X., Telmer, P., Marsolais, F., and Dhaubhadel, S. (2010). A single-repeat MYB transcription factor, GmMYB176, regulates *CHS8* gene expression and affects isoflavonoid biosynthesis in soybean. *Plant J.* 62, 1019–1034. doi: 10.1111/j.1365-313X.2010.04214.x

Zhang, Y., Sun, Y., Sun, J., Feng, H., and Wang, Y. (2019). Identification and validation of major and minor QTLs controlling seed coat color in *Brassica rapa* L. *Breed. Sci.* 69, 47–54. doi: 10.1270/jsbbs.18108

Zhang, Z. S., Xiao, Y. H., Luo, M., Li, X. B., Luo, X. Y., Hou, L., et al. (2005). Construction of a genetic linkage map and QTL analysis of fiber-related traits in upland cotton (*Gossypium hirsutum* L.). *Euphytica* 144, 91–99. doi: 10.1007/s10681-005-4629-x

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi: 10.1038/nbt.3096

Zhuang, Y., Li, X., Hu, J., Xu, R., and Zhang, D. (2022). Expanding the gene pool for soybean improvement with its wild relatives. *aBIOTECH* 22, 115–125. doi: 10.1007/s42994-022-00072-7

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 28 September 2022
DOI 10.3389/fpls.2022.997860

OPEN ACCESS

EDITED BY
Jinyoung Y. Barnaby,
United States Department of Agriculture (USDA), United States

REVIEWED BY
João Ricardo Bachega Feijó Rosa,
Federal University of Viçosa, Brazil
Tudor Borza,
Dalhousie University, Canada

*CORRESPONDENCE
Johanna Osterman
johanna.osterman@slu.se

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

RECEIVED 19 July 2022
ACCEPTED 07 September 2022
PUBLISHED 28 September 2022

CITATION
Osterman J, Hammenhag C, Ortiz R and Geleta M (2022) Discovering candidate SNPs for resilience breeding of red clover. *Front. Plant Sci.* 13:997860. doi: 10.3389/fpls.2022.997860

Discovering candidate SNPs for resilience breeding of red clover

Johanna Osterman*, Cecilia Hammenhag, Rodomiro Ortiz and Mulatu Geleta

Department of Plant Breeding, Swedish University of Agricultural Sciences, Lomma, Sweden

Red clover is a highly valuable crop for the ruminant industry in the temperate regions worldwide. It also provides multiple environmental services, such as contribution to increased soil fertility and reduced soil erosion. This study used 661 single nucleotide polymorphism (SNP) markers via targeted sequencing using seqSNP, to describe genetic diversity and population structure in 382 red clover accessions. The accessions were selected from NordGen representing red clover germplasm from Norway, Sweden, Finland and Denmark as well as from Lantmännen, a Swedish seed company. Each accession was represented by 10 individuals, which was sequenced as a pool. The mean Nei's standard genetic distance between the accessions and genetic variation within accessions were 0.032 and 0.18, respectively. The majority of the accessions had negative Tajima's D, suggesting that they contain significant proportions of rare alleles. A pairwise F_{ST} revealed high genetic similarity between the different cultivated types, while the wild populations were divergent. Unlike wild populations, which exhibited genetic differentiation, there was no clear differentiation among all cultivated types. A principal coordinate analysis revealed that the first principal coordinate, distinguished most of the wild populations from the cultivated types, in agreement with the results obtained using a discriminant analysis of principal components and cluster analysis. Accessions of wild populations and landraces collected from southern and central Scandinavia showed a higher genetic similarity to Lantmännen accessios. It is therefore possible to link the diversity of the environments where wild populations were collected to the genetic diversity of the cultivated and wild gene pools. Additionally, least absolute shrinkage and selection operator (LASSO) models revealed associations between variation in temperature and precipitation and SNPs within genes controlling stomatal opening. Temperature was also related to kinase proteins, which are known to regulate plant response to temperature stress. Furthermore, the variation between wild populations and cultivars was correlated with SNPs within genes regulating root development. Overall, this study comprehensively investigated Nordic European red clover germplasm, and the results provide forage breeders with valuable information for further selection and development of red clover cultivars.

KEYWORDS

pool-seq, phylogenetic tree, DAPC, LASSO, bioclimatic variables, red clover

Introduction

Red clover is a perennial forage legume that grows in temperate regions worldwide. Due to its high protein content, it is considered an important crop for the ruminant industry (Smith et al., 1985; Taylor and Quesenberry, 1996a). In addition to its great nutritional value, red clover provides several important ecosystem services. Due to its symbiotic relationship with nitrogen-fixing bacteria (Sturz et al., 1997; Thilakarathna et al., 2017), red clover increases soil fertility. Compared to alfalfa and white clover, which have similar symbiotic relationships with the *Rhizobium* bacteria, red clover is more efficient at nitrogen fixation (Dhamala et al., 2017). As a perennial crop, red clover also contributes to soil carbon sequestration, reduces soil erosion during the winter, and suppresses weeds (McKenna et al., 2018). However, persistence is generally low in red clover, which adversely affects its overall performance as a forage crop (Taylor and Quesenberry, 1996b). Red clover is a diploid species ($2n = 2x = 14$); however, tetraploid ($2n = 4x = 28$) cultivars have been developed through chromosome doubling techniques (Taylor and Quesenberry, 1996c). The tetraploid red clover cultivars generally have a higher green biomass yield as well as a higher persistence and resilience than the diploids (Öhberg, 2008). However, their seed yield is generally lower than that of the diploids because of their flower anatomy and a higher rate of embryo abortion (Amdahl et al., 2016).

The development of modern DNA marker-based plant breeding techniques for red clover is lagging behind, despite its economic and ecological significance, although it has been picking up pace in recent years. For instance, the publication of its reference genome (De Vega et al., 2015) has facilitated the discovery of quantitative trait loci (QTL) for various traits, and the "mining" of single nucleotide polymorphism (SNP) markers (Herrmann et al., 2008; Ergon et al., 2019; Li et al., 2019). In two recent papers on population genetics, SNPs were used to assess the population structure in individually genotyped red clover (Jones et al., 2020; Osterman et al., 2021). Jones et al. studied 75 accessions from Europe, Asia, and Iberia where 70 were wild populations and five were commercially available breeding populations. They found that the population structure of red clover is highly correlated with its geographical location and associated climatic conditions. Osterman et al. focused more on the genetic differences between accessions representing different populations and found that, for instance, wild populations were clearly differentiated from cultivated populations. Both studies noted the effect of the outcrossing nature of red clover in the overall higher heterozygosity which decreases the levels of genetic differentiation. Since red clover is a strictly outcrossing species, genetic research should ideally be performed at a population level. Currently, it is quite expensive to sequence an adequate number of individuals within each population for a comprehensive genetic analysis of multiple populations. With a method that is generally referred to as Pool-seq (Schlötterer et al., 2014), individuals can be pooled and sequenced simultaneously using different next-generation sequencing (NGS) methods. SeqSNP is a targeted genotype by sequencing method for genotyping known SNPs, which is also amenable to *de novo* discovery of SNPs located close to the target SNP positions (Osterman et al., 2021). Hence, Pool-seq via SeqSNP is an NGS method in which individuals are sampled, pooled, and sequenced together, targeting known SNP loci. The target SNPs can be selected from the available SNP databases or developed through allele-mining approaches based on existing genomic resources.

SNP markers are codominant single nucleotide markers that have been widely used in various applications, including genomic selection (GS; Heffner et al., 2009) and marker-assisted selection (MAS; Lande and Thompson, 1990). Compared to the phenotype-based selection, these two breeding methods are quicker and can facilitate the development of high-yielding cultivars that are resilient and nutritious within a shorter period of time. Gene-specific SNP markers are preferred over SNPs in other genomic regions since they are more likely to be associated with genes that regulate desirable traits (Poczai et al., 2013). Hence, genes that are highly desirable from the viewpoint of plant breeding can be targeted for genetic diversity analyses. This will enable the determination of suitable genetic resources that could be used in plant breeding programs. Because genetic similarity between populations might reflect phenotypic similarity, gene-specific markers could provide crucial insights into the differentiation of populations in terms of traits, such as growth and development.

Due to its proximity to the North Pole and the effects of the passing Gulf Stream, the Nordic Region of Europe has highly variable weather with large differences in day length over seasons despite its geographically small area. Consequently, the region requires unique crop cultivation conditions, and the key to crop persistence could be found in its wild relatives. In this regard, genetic analyses of both wild and cultivated gene pools could link the breeding material used by Scandinavian breeding enterprises to resilient wild populations.

The purpose of this study was to compare and examine the genetic resources of red clover available in northern Europe by targeting its cultivated and wild gene pools that represent the Nordic countries. Here, SeqSNP was used to target SNPs within genes that influence growth and development, as well as disease resistance. Moreover, population genetic analyses were carried out in order to determine the correlation between genetic differences among wild populations and bioclimatic variables at the original collection sites.

Material and methods

Selecting germplasm and planting

For this study, 382 accessions of red clover were used that originate from different parts of the Nordic Region of Europe

(Supplementary Table 1). Of these, 294 accessions were obtained from NordGen (a regional genetic resources center for the Nordic countries) and 88 accessions from Lantmännen Seed (a plant breeding and agricultural seed company based in Sweden). The NordGen accessions were selected based on their passport data to represent a variety of available germplasm (cultivars, breeding populations, landraces, and wild populations) representing most of the Nordic region of Europe (i.e., Sweden, Norway, Finland, and Denmark). One Russian accession was also included as it was located at the Finnish border. The Lantmännen varieties include cultivars and synthetic populations from the Scandinavian forage breeding programs (Lantmännen, Sweden; Graminor, Norway; and DLF, Denmark), and hence reflect the variety of cultivated red clover available in northern Europe. These accessions include both diploid and tetraploid types that are categorized either as late or middle-late based on their maturation period.

The accessions were planted and grown for two weeks in a greenhouse at the Swedish University of Agricultural Sciences (SLU, Alnarp, Sweden), as described in Osterman et al. (2021). The BioArk leaf collection kit (LGC Biosearch Technologies) was used to collect ten 6 mm leaf discs (1 leaf disc/plant). One pool of leaf tissue representing ten plants was sampled for each accession separately. DNA extraction and genotyping were conducted at LGC Biosearch Technologies (Berlin, Germany), as described in Osterman et al. (2021).

Genotyping and variant calling

For genotyping, SeqSNP was used with a set of 400 target SNPs developed by Osterman et al. (2021). The SNPs were identified from publicly available red clover genomic resources by targeting coding sequences of genes known to be involved in growth and development as well as in the response to biotic and abiotic stresses. In addition to genotyping the target SNPs, SeqSNP was used to discover novel SNPs in the regions surrounding the target SNPs. In this analysis, 2 × 150 bp reads were used as a sequencing mode. The sequencing depth was set to 50 million read pairs (15 GB raw data) per sample to ensure sufficient coverage of each genotype in each pool thus adjusting the sequencing depth to ×501. SNP calling was performed by aligning the quality trimmed reads to the reference genome using Bowtie2 v2.2.3. For variant discovery, Freebayes2 was used with ploidy set to diploid and minor allele frequency set to 1%. To exclude calls due to sequencing error, allele counts below eight were set to zero as per the recommendations of LGC Biosearch Technologies where genotyping was conducted. The allele frequency of each accession at each locus was calculated based on the read counts.

For determining the validity of converting the read counts of pool-seq into allele frequencies for data analysis, five randomly selected accessions were genotyped at both the individual and pool levels. Following this, the read counts of each pooled sample were converted to allele frequencies. A subsequent step involved converting the genotypic data of the five individuals of each accession into allele frequency data for that particular accession. This was followed by correlation analysis between the allele frequencies obtained from individual genotype sequencing and PoolSeq, which revealed a highly significant correlation ($r > 0.95$; $P < 0.001$) between them. Hence, SeqSNP is a highly reliable method to generate data for allele frequency-based data analyses.

Among the 400 target SNP loci genotyped, 5.5% were mono-allelic, 86% were bi-allelic, 7.5% were tri-allelic, and 0.75% were tetra-allelic across the 382 accessions studied. The remaining 0.25% were INDELs (insertion or deletion of a nucleotide). The *de novo* SNP and INDEL calling generated 347 SNPs and 16 INDELs. Among the 347 novel SNPs, 91% were bi-allelic, 8% were tri-allelic, and 1% were tetra-allelic. Due to the complexity of analyzing pooled sequencing data and a mixture of diploid and tetraploid accessions, only polymorphic bi-allelic markers were used. Additionally, tetraploids were treated as diploids as described in Osterman et al. (2021). Overall, 344 originally targeted and 317 *de novo* discovered bi-allelic SNPs (661 SNPs in total), all with minor allele frequency of 5% or above, were used for data analyses in this study.

Genetic parameter estimation

Tajima's D was estimated using PoPoolation (Kofler et al. 2011b) based on the quality- trimmed reads combined in a sync file using the respective reference sequences to map the reads. The allele counts from Freebayes2 were imported into R (R Core Team, 2013) and the expected heterozygosity for each population (H_S) was calculated using the adegenet package (Jombart, 2008). Using the poolfstat package (Gautier et al., 2022) in R, pairwise F_{ST} was calculated for each pair of SNPs as well as for each pair of accessions. After grouping the accessions according to their origins or types an additional pairwise F_{ST} analysis was performed. Nei's standard genetic distance between populations and between groups (as for the F_{ST} analysis) was calculated using adegenet package. Additionally, Mantel's randomized test comparing Nei's standard genetic distance with the geographic coordinates of the germplasm collecting sites of the wild populations was performed to determine whether isolation by distance (IBD) has a significant effect on the genetic differences between the accessions.

Determining population structure via clustering

Both principal component analysis (PCA) and principal coordinate analysis (PCoA) were used to determine the genetic relationships between the accessions. The PCA was conducted

using the pcadapt package (Luu et al., 2017), and SNPs that were most associated with the variation described in the first two principal components were extracted for further analysis. The PCoA was performed using the stats package in R (R Core Team, 2013) based on the Nei's genetic distance. The Nei's genetic distance based relationship between the accessions was further analyzed using ComplexHeatmaps (Gu et al., 2016), which generates heatmaps. These analyses enabled the comparison of the accessions both individually as well as collectively based on their origins and types.

The Nei's genetic distance based relationship between the accessions was further investigated through neighbor-joining (NJ) cluster analysis. The NJ tree was built using the ape package (Paradis et al., 2004) and visualized using the ggtree package (Yu et al., 2017). Incorporating bioclimatic variables to the NordGen accessions of wild populations and maturity types to the Lantmännen accessions into the analyses was made possible using the ggtree package. A map of collection coordinates for landraces, some cultivars and breeding populations as well as wild populations provided by NordGen was constructed using the rnaturalearth package, which uses maps from Natural Earth, in R.

A discriminant analysis of principal components (DAPC) was performed using adegenet with the method described by Jombart et al. (2010); Jombart and Collins (2015) on the allele frequencies. The find.clusters function was used to find the most optimal number of clusters based on the BIC score, and the cluster solution with the lowest BIC score was chosen. The xval function with a test set of 90% with 30 repetitions was used to find the appropriate number of principal components (PCs). This resulted in a five-cluster solution involving 150 PCs.

Environmental data for NJ tree and LASSO models

Bioclimatic variables were retrieved from WorldClim (Fick and Hijmans, 2017) with a spatial resolution of 30 seconds (~ 1 km^2) and imported via the raster (Hijmans et al., 2012) package in R. The coordinates of the germplasm collecting sites of the wild populations were used to extract environmental data with interpolation, hence minimizing the effect of potential recording errors. The bioclimatic variables were evaluated based on how they vary within Scandinavia. Most of the precipitation variables were similar across the sampling sites, and therefore they were excluded. The final set of bioclimatic variables from WorldClim include annual mean temperature and precipitation, as well as isothermal and precipitation seasonality. Annual snow coverage was estimated using snow thickness data retrieved from Climate Data Store (CDS) for months with snow (September to June) from 1980 to 2000. The mean snow thickness for each month at a sampling site during the years 1980 to 2000 was calculated and plotted with months on the x-axis and mean snow thickness on the y-axis. The area under the curve (AUC) of the yearly trend was calculated using the AUC function from the DescTools package (Andri, 2021) in R. The AUC value was chosen to represent mean snow coverage for each collection site.

SNPs with significant contribution to the variation explained by PCA

The pcadapt package (Luu et al., 2017) was used to perform a PCA with the Mahalanobis' method as the function argument. Thus, the Mahalanobis' distance was used to measure the extent to which each SNP is related to the first, in this case, two PCs. A χ^2-test was performed on the SNPs Mahalanobis distance to find those SNPs with significant contribution to the population structure, and the Benjamini–Hochberg correction was applied to control false positive discovery. The significant SNPs were then validated for their biological roles using gene ontology (GO) enrichment to find possible molecular functions differentiating the populations in the PCA clusters.

LASSO models

The least absolute shrinkage and selection operator (LASSO) regression model was used to connect environmental variables to the allele frequencies. As the number of SNPs exceeded the number of populations in this study, LASSO was considered an appropriate model due to the application of penalization and feature reduction. Among the accessions studied, only those representing wild populations were used for the LASSO models. This is because they could be considered to have adapted to the environments of the sampling sites. Considering the number of SNPs available for the data analysis was high (661), a method for selecting only the most relevant ones was devised to increase the biological aspect of model training. The goal of this analysis was to find the SNPs that genetically differentiated two populations. Thus, the F_{ST} values of all 661 SNPs between every pair of populations were calculated and the top 1% values of each pairwise calculation was selected. A final set of 430 SNPs with high F_{ST} values was selected for the LASSO models. The caret package (Kuhn, 2008) in R was used to train and select the LASSO model. A leave one out cross validation (*LOOCV*) approach was used to train a regression model (*glmnet*) and the mean average error (*MAE*) was computed for model selection. The MAE and root mean square error (RMSE) were compared to the variance of each climate variable to validate the final model; and the error was smaller than the standard deviation of the input for all models. To validate the models further, a linear regression analysis was performed on the bioclimatic variables without including the SNP data. In cases where the LASSO model had a lower RMSE than the linear model, the SNPs were considered to have an effect.

Validating selected SNPs in terms of biological functions

The analysis of SNP effect on population differentiation in the PCA and LASSO models resulted in nine sets of SNPs, two from the PCA and seven from the LASSO models. A gene ontology (GO) enrichment analysis was carried out on each set of SNPs to find any biological function underlying the population differentiation. The GO analysis was performed with the workbench at dicots Plaza 5.0 (Van Bel et al., 2021) where the correct names of the genes containing the SNPs were determined via the integrated BLAST function. Then enrichment was performed using all red clover genes as background with p-values showing significant enrichment adjusted following Bonferroni's correction.

Results

The SeqSNP-based sequencing of the 382 red clover accessions resulted in 661 bi-allelic SNP markers, which were then used for population genetics analysis of the accessions (Supplementary Table S2; Osterman et al., 2021). Additionally, 49 tri-allelic, four tetra-allelic and 17 INDELs were identified across the 400 target SNP loci, and 357 SNP loci were discovered *de novo*. Of the 317 *de novo* discovered bi-allelic SNPs, 292 were reported in Osterman et al. (2021) whereas the remaining 25 were specific to this study. It is evident from the number of *de novo* SNPs discovered in this study, compared to that of Osterman et al. (2021), the number of accessions studied had an effect on the number of novel SNPs discovered. Only bi-allelic SNPs were used for the data analyses for the sake of simplicity. At each of the 661 bi-allelic SNP loci, the allele frequency was calculated based on the read counts obtained from the sequencing. The read counts across the 661 bi-allelic SNPs ranged from eight to 4320. Although the range of the allele counts is large, there was no need to scale the frequencies since they were calculated independently for each locus of each accession.

Genetic variation within and among groups

For data analysis, the accessions were grouped based on their origins and population types. The grouping results in nine origin-based groups (Denmark, Finland, Graminor, Lantmännen, Norway, Sweden, DLF, Local population, and Russia) and eight population type-based groups (Breeding population, Cultivar, Diploid, Graminor, Landrace, Tetraploid, Unknown, and Wild Population). Because DLF, Local population, and Russia (among the origin-based groups) and Graminor and Unknown (among the population type-based groups) had only one accession each, they were excluded from some analyses.

The study revealed low genetic diversity and population structure considering the median and mean values of Nei's standard genetic distance and F_{ST} of each group (Table 1, Figure 1). Additionally, the results show a difference in the amount of rare alleles present within groups where wild populations had larger levels of rare alleles than cultivated accessions (Tajima's D in Table 1 and Figure 1). All cultivated groups (breeding populations, cultivars, diploids and tetraploids) had negative F_{ST} mean values. Both negative and zero F_{ST} values indicate lack of genetic variation distinct to each of the populations compared. Only wild populations had a positive mean F_{ST} value, hence, it is the only group (among population types) with any population structure. In the case of origin-based groups, the mean F_{ST} values were negative for Lantmännen, Graminor and Denmark, zero for Finland, and positive for Norway and Sweden. Apparently, the F_{ST} values of the different population type-based groups and origin-based groups were related due to the accessions they shared. The majority of the landrace accessions belong to Finland, and consequently the mean F_{ST} values for both groups were zero. Similarly, most of the accessions from Denmark were cultivars, and hence both Denmark (origin) and cultivars (population type) had a negative mean F_{ST} value. The mean F_{ST} for Sweden and Norway was higher, as they were mainly comprised of wild populations.

The pairwise F_{ST} between groups showed high genetic similarity between the cultivated types (Figure 2A) while the wild population group was divergent from the rest. Among the origin-based groups, Sweden, Norway, and Russia (which are dominated by wild populations) showed significant genetic differentiation from the other origin-based groups (forming a separate cluster in Figure 2B). This suggests a significant difference in allelic states between accessions from these countries and those from the other origins. Further, cultivated types as well as origin-based groups that are largely composed of cultivated types did not appear to have a clear population structure within or between them.

The mean values of Nei's standard genetic distance within the different groups were quite similar (Table 1). Hence, it was further investigated at a population level to illustrate groups of populations with high genetic similarity and had similar genetic relationships with the remaining populations. Only a few groups could be identified in the present study (marked by the blue rectangles in Figures 3A–C). There was a clear separation between clusters containing populations with a relatively high genetic distance (wild populations) and populations with low genetic distance (cultivars, Figure 3A). When the wild populations were separately analyzed, two clusters were identified, where one contained mainly the Swedish and Norwegian populations while the other contained mostly Swedish but also Finnish and Norwegian populations (Figure 3B). The separate analyses of the Lantmännen accessions

10.3389/fpls.2022.997860

TABLE 1 The first column indicates the number of samples in different groups of red clover populations grouped according to their origin or population type.

Grouped by origin	N° samples	H_s	Nei	F_{st}	Tajima's D	B	C	D	G	L	T	U	W	
Denmark[a]	35	0.19	0.02	-0.01	-0.5	3	86						3	
Finland[a]	72	0.18	0.03	0	-0.03		4			88			15	
Graminor[a]	6	0.19	0.02	-0.03	0.02			33	17	50				
Lantmännen[a]	81	0.19	0.02	-0.02	-0.01			44			56			
Norway[a]	92	0.18	0.03	0.03	-0.04	5	1			4			89	
Sweden[a]	95	0.18	0.03	0.03	-0.05	1	7			9		1	81	
DLF[a]	1	0.20	-	-	-0.02						100			
Local population[a]	1	0.20	–	–	0.06		100							
Russia[a]	1	0.16	-	-	0								100	
Grouped by type						**Da**	**F**	**G**	**La**	**N**	**S**	**Df**	**Lo**	**R**
Breeding population[b]	10	0.19	0.02	-0.02	0.07	40				50	10			
Cultivar[b]	41	0.19	0.02	-0.02	-0.05	74	7			2	17			
Diploid[b]	43	0.19	0.03	-0.01	-0.01			5	93				2	
Graminor[b]	1	0.19	–	–	0.01			100						
Landrace[b]	71	0.19	0.03	0	-0.02		81			6	13			
Tetraploid[b]	45	0.20	0.02	-0.04	-0.01	91		7				2		
Unknown[b]	1	0.17	-	-	0.04						100			
Wild population[b]	172	0.18	0.04	0.04	-0.05	0.5	6			48	45			0.5

[a]= a group of accessions belonging to geographic origin; [b]= a group of accessions belonging to population type; H_s, mean expected heterozygosity; Nei, Nei's standard genetic distance; F_{st}, mean fixation index; Tajima's D, Tajima's population genetic test statistic.
The second to fifth column is the mean of the genetic parameter for the group. The last five columns refer to the composition of each group. When grouped by origin it is the percentage of population types, **B**reeding population, **C**ultivar, **D**iploid, **G**raminor, **L**andrace, **T**etraploid, **U**nknown and **W**ild population. When grouped by type the columns refer to the percentage of **De**nmark, **F**inland, **G**raminor, **La**ntmännen, **N**orway, **S**weden, **DL**F, **Lo**cal population or **R**ussian federation.

representing the cultivated gene pool revealed similarly low genetic distance between the accessions, and no clearly defined clusters were found (Figure 3C). Similar results were obtained with the cultivated accessions of NordGen, with the exception of a small cluster formed by the Finish landrace accessions (Figure 3D).

Cluster analysis via PCoA, DAPC and neighbor joining tree

A principal coordinate analysis conducted based on Nei's standard genetic distance showed that the first principal coordinate (PCo1), which accounted for 30.8% of the total variation distinguished most of the wild populations from the cultivated ones (Figure 4A). It was also shown that the landrace populations were placed between the wild and cultivated populations along the PCo1. The second principal coordinate (PCo2), which accounted for 10.9% of the total variation, distinguished wild populations and one landrace population originating from Norway from a group containing wild populations from Sweden, Finland, and Russia, as well as landrace populations from Finland. The results clearly showed that the major contributors to the variation displayed in the first two principal coordinates are wild populations. Thus, to get a better understanding of the main clusters, the 382 accessions were divided into subsets. In the wild population subset, the pattern persisted as expected and the cumulative variance described by the first two PCos decreased only slightly (from 41.7% to 39.1%; Figure 4B).

When the Lantmännen accessions were separately analyzed, the PCoA showed a cumulative variance of 23.9% in the first two PCos (Figure 4C). However, the scatter plot showed no clear partition between diploid and tetraploid accessions in both the PCo1 and PCo2. The separate PCoA of the NordGen accessions revealed a major separation of the landrace accessions from cultivars and breeding populations along the first PCo, which explained 24.8% of the total variation. Furthermore, it showed a separation of populations from Denmark and Finland (Figure 4D). The second PCo described far less variation (8.3%) and did not show a clear separation between any of the different groups.

A DAPC on the 382 accessions explained 79.2% of the total variance and revealed five clusters (Figure 5 and Supplementary Table S1). Clusters 3 and 5 mainly comprised the cultivated types as well as some wild accessions. The major source of variation for the differentiation between clusters 3 and 5 appears to be the accessions' countries of origin, especially Denmark versus Finland. Clusters 1, 2 and 4 differentiated the wild populations from the cultivated types although some landraces were contained in Clusters 1 and 4. Clusters 2 and 4 are

10.3389/fpls.2022.997860

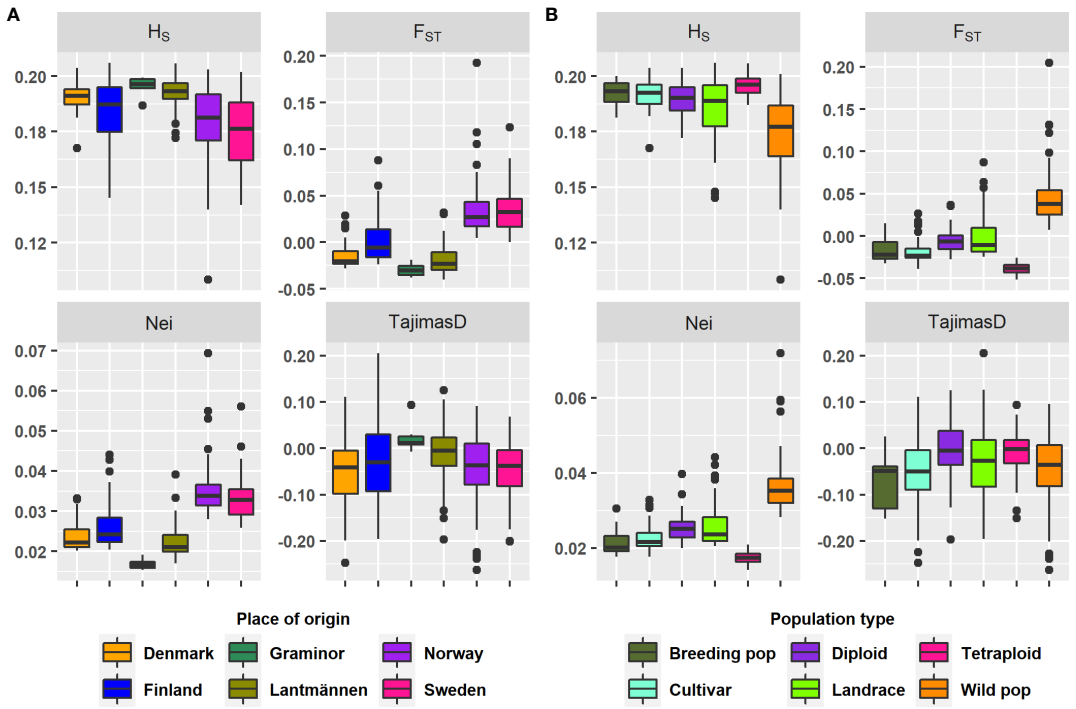


FIGURE 1
A box plot depicting the range and median for the genetic parameters on each group according to **(A)** Origin and **(B)** Type. The genetic parameters were H_S, mean expected heterozygosity; Nei, Nei's standard genetic distance; F_{ST}, mean fixation index; Tajima's D, Tajima's population genetic test statistic.

dominated by wild populations from Sweden and Norway, respectively, while Cluster 1 comprised of wild populations and landraces from Finland and Sweden. This clustering follows the map Nordic Region of Europe from east to west.

Each accession has been assigned to a cluster based on a membership probability, which can be plotted in the same way as the commonly used software STRUCTURE Following the instructions provided in Jombart and Collins (2015) (Figure 6). The membership probabilities were high with some overlaps between cluster 1 and 2, 1 and 3 as well as 1 and 4. Here, the differentiation between Finnish and Danish populations in clusters 1 and 4 is more prominent. It is again shown by the membership of wild populations in all clusters, that the wild populations contain a high genetic variance.

The differentiation between the cultivated and wild populations was again observed in a neighbor-joining cluster analysis based on Nei's standard genetic distance where the 382 accessions formed four major clusters (Figure 7 and Supplementary Table S1). Cluster-1 contained the majority of the wild populations, some cultivated types from both NordGen and the breeding companies. The NordGen cultivated types comprised three breeding populations from Norway, Denmark and Finland and three cultivars from Finland and Sweden. Whereas the Lantmännen cultivated types include one Graminor population and three diploid cultivars from Lantmännen. Cluster-2 and cluster-3 contained the majority of the Lantmännen accessions. Interestingly, cluster-4 contained almost exclusively Finnish accessions with the exception of one diploid cultivar and tetraploid cultivar from Lantmännen and one Norwegian wild population. Wild populations and landraces in cluster-2, cluster-3 and cluster-4 originated from along the coast or near a lake in southern to central Scandinavia. The Mantel test, which compared geographical distances with Nei's genetic distance, revealed that isolation by distance is evident (Supplementary Figure 1), indicating that environmental variance could be linked to genetic variation.

The bioclimatic data from WorldClim successfully described the local environment at each of the wild populations' sampling sites, demonstrating the variation of climate factors in the region. The five bioclimatic variables, shown in columns 4 to 8 in the heatmap of Figure 7, describe the average environment for each sampling site of the wild populations. They show, for example, a connection between lower annual mean temperatures and higher mean snow coverage. The highest snow coverage was recorded in the most northern geographical locations. Additionally, there were several geographical locations with high annual mean temperatures as well as high snow coverage. Similarly, there were multiple sites with high isothermality, i.e. a large difference in day to night

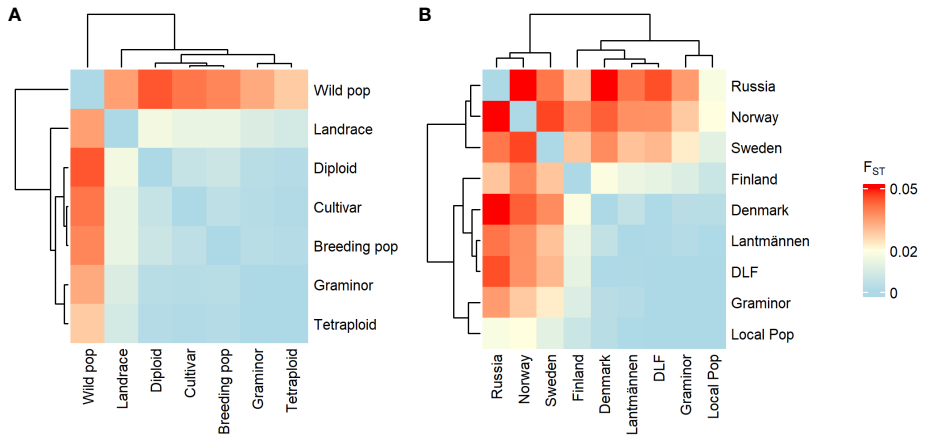


FIGURE 2
Heatmap depicting the pairwise F_{ST} values between groups of red clover populations based on population type **(A)** or origin **(B)**.

temperatures between summer and winter. Wild red clover from these locations would have developed resilience to the harsh winter conditions, to which cultivated red clover is highly susceptible.

Interestingly, despite the fact that the wild populations with close genetic relationship to Lantmännen material do not span the entire geographical area of interest, they still represent most of the climatic conditions. Nevertheless, the main climatic conditions associated with the Lantmännen breeding materials were warmer, steady temperatures and low snow coverage. Additionally, populations with similar maturation periods clustered together in cluster-2 and cluster-3. All populations in cluster-4 with a known maturation period, except one, were late maturing. Contrary to the hypothesis, that late maturing would have larger similarities with northern wild populations, nevertheless there was no clear distinction between the maturity groups.

LASSO models and GO analysis

In the present study, the parameter selection feature of LASSO models was used to estimate the SNPs that were the most informative in predicting the values of bioclimatic variables. A model with a root mean square error (RMSE) smaller than the standard deviation (SD) indicates that there is a predictive effect of the feature (SNP). In other words, there is an effect of the selected markers on the predictive ability of the model. All LASSO models had RMSE smaller or about equal to their respective SD (Table 2). Thus, all models had good predictive ability except for mean snow coverage and isothermality. The goodness of fit of the model was further confirmed by conducting a simple linear regression analysis without using the SNP information and comparing the RMSE. The RMSE of the linear regression was closer to the SD than the RMSE of the LASSO for all models except isothermality and snow coverage, thereby confirming the effect of the SNPs in the model's prediction.

The gene ontology (GO) enrichment analysis was used to validate the model results, in terms of biological functions. The reference gene-coding sequences, of the SNPs selected by the LASSO models were imported into the online tool Plaza workbench via BALST to find the corresponding genes. The models in which SNPs generated a GO enrichment were annual mean temperature, annual precipitation, and annual temperature range (Supplementary Table S3). Interestingly, the set of genes from both the annual precipitation and annual temperature models showed enrichment for genes regulating stomatal opening (Table 2 and Supplementary Table S3). The stomata are known to be involved in the plants regulation of water, oxygen and carbon dioxide, functions that are relevant to changes in temperature and precipitation (Waggoner and Zelitch, 1965; Honour et al., 1995). Furthermore, there was an enrichment of genes coding for kinase binding proteins in the annual mean temperature model (Table 2; Supplementary Table S3). Kinases are a group of enzymes that via post-translational modification plays an important role in plant growth and development. Some kinases are involved in the plant response to changes in both mild and extreme temperatures (Praat et al., 2021). Analogs of the three genes detected in the GO enrichment analysis were, via experimental evidence, connected to heat response (Larkindale et al., 2005; Wu et al., 2014) and to post translational modifications as response to external factors (Colby et al., 2006) in tomato and *Arabidopsis*.

The GO analysis of SNPs from PCA

From the PCA analysis, using the R software's pcadapt package, the SNPs that significantly contributed to the

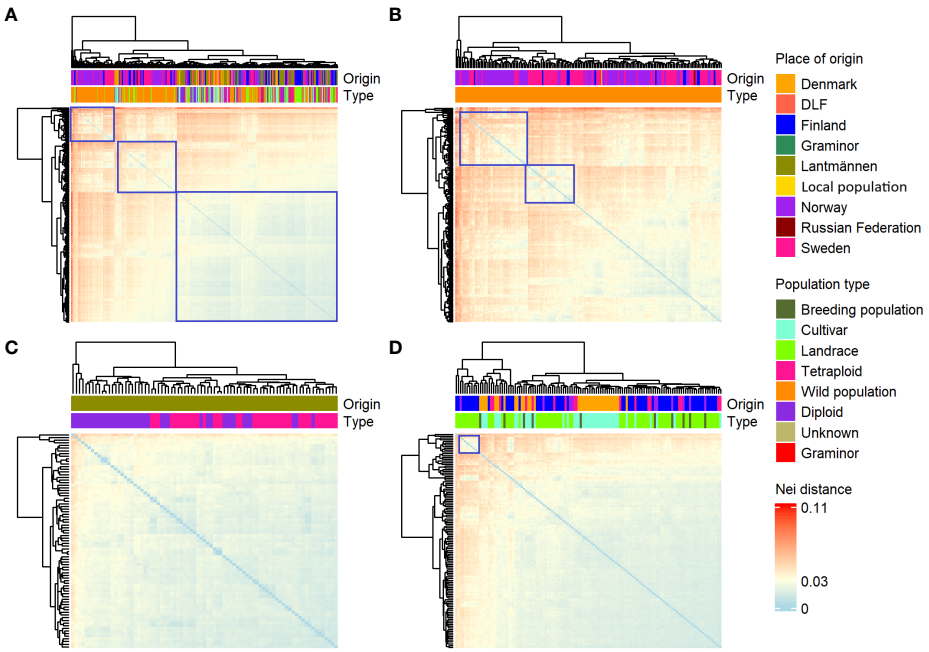


FIGURE 3
A heatmap depicting Nei's standard genetic distances between each pair of populations. The colors indicate high (red), intermediate (yellow) or low (blue) genetic distances. The accessions were clustered according to their pairwise genetic similarities. The accessions included were **(A)** all 382; **(B)** WildOnly those that are wild; **(C)** PopulationsOnly those from Lantmännen; and **(D)** Cultivarsonly cultivars and landraces from NordGen.

clustering of the 382 accessions in the first two PCs were located within the coding regions of 22 and 38 genes, respectively (Supplementary Table S3). A GO enrichment analysis of the 22 genes from PC1 revealed that 10.5% of the genes were enriched for two biological processes, namely, specification of plant organ axis polarity and regulation of root morphogenesis. However, no GO enrichment was observed for the 38 genes from PC2.

Discussion

This study revealed the genetic variation of 382 red clover accessions, including wild and cultivated types representing the red clover gene pool in the Nordic Region of Europe. Among the red clover accessions studied, 45 were known to be tetraploids. However, in order to facilitate the comparison with the diploid populations used in this study they were treated as diploids, following the explanation provided in Osterman et al. (2021). Diploidizing tetraploids is commonly employed to reduce complexity in data analysis and has been implemented in potato for population structure analysis using STRUCTURE and other software developed for diploids (Hirsch et al., 2013; Pandey et al., 2021; Selga, 2022). The genotyping was conducted using a pool-seq approach of the SeqSNP sequencing assay used by Osterman et al. (2021) that targeted genes known to be involved in growth and development as well as stress response. In total, 661 polymorphic, bi-allelic SNPs were detected in the targeted protein coding sequences across the 382 populations, demonstrating the potential of SeqSNP for sequencing the target SNPs as well as for *de novo* SNP discovery. Nevertheless, it should be noted that the novel SNPs identified were within 75 bp of the target SNPs. Therefore, SeqSNP is useful when specific coding regions are targeted, but may not be suitable for the identification of novel SNPs in larger regions, such as quantitative trait loci (QTL) and uncharacterized gene sequences.

The benefit of using pool-seq methods as opposed to individual sequencing methods is the number of populations that can be analyzed. With a pool of 10 individuals, pool-seq can analyze 10 times as many populations as individual genotype analysis for the same cost, assuming their sequencing depth is the same. The main challenge of pool-seq is the data analysis, as the representation of the sampled individuals within a pool can be uneven. The selection of reading depth relative to the number of individuals in each pool is very important. In the present study, using read counts from ten individuals at a sequencing depth of x501 was deemed appropriate, given the result of our in-house analysis that compared data generated through pool-seq and individual genotype sequencing.

The genetic parameters estimated from ten individuals per pool were comparable with the results reported by Jones et al. (2020)

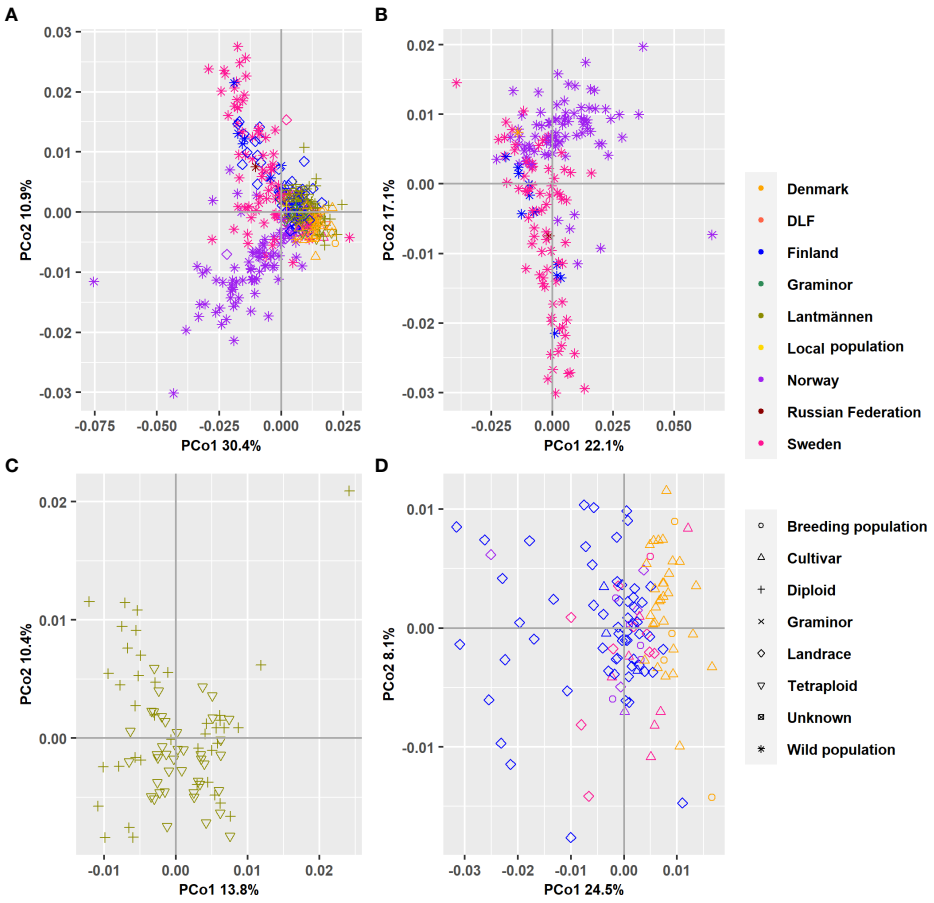


FIGURE 4
A bi-plot of the principal coordinate analysis (PCoA) showing the variation explained by the first two principal components. **(A)** All 382 populations' analyzed together and separate analysis when the accessions have been grouped according to wild populations **(B)** Lantmännen populations **(C)** or Landraces and cultivars held at NordGen **(D)**.

and Osterman et al. (2021). Various bioinformatics software packages have been developed to analyze pool-seq data, including BayPass (Gautier, 2015), PoPoolation (Kofler et al., 2011a; Kofler et al., 2011b), and SelEsim (Vitalis et al., 2014). Pool-seq based study on red clover using BayPass has previously been done to detect genomic signatures of herbicide resistance (Benevenuto et al., 2019). The study used pools of 20 to 40 individuals from 10 populations and data analysis was performed using BayPass, which uses read counts and a hierarchal Bayesian model to estimate genetic variance/covariance and outlier loci. However, in the present study, the number of populations (382) relative to the number of SNPs used was too large to apply BayPass.

The discovered genetic variation differentiated the groups of accessions to different extents based on their population type or origin. The major trend was a separation of the wild populations from the cultivars, with the landraces being represented within both groups. The genetic distance between the cultivated populations was low, and there was no clear separation between them based on their origins. This could be partly due to the strict outcrossing nature of the crop that facilitated a high rate of gene flow, resulting in high heterozygosity with reduced differences in allele frequency between the populations. A high level of heterozygosity has been previously reported in red clover populations analyzed at individual genotypes level, which was attributed to its strict outcrossing reproductive system (Jones et al., 2020; Osterman et al., 2021). Contrary to this, there was a pattern of population structure between the different groups of wild populations. The lower rate of gene flow between the wild populations is probably due to the low level of migration as well as the consequences of geographical distance and terrain.

This study was designed to identify informative SNPs from the perspective of red clover breeding and to generate knowledge regarding the extent to which the genetic material used for breeding reflects available genetic resources in red clover. These objectives are clearly reflected in the NJ tree (Figure 7) as well as

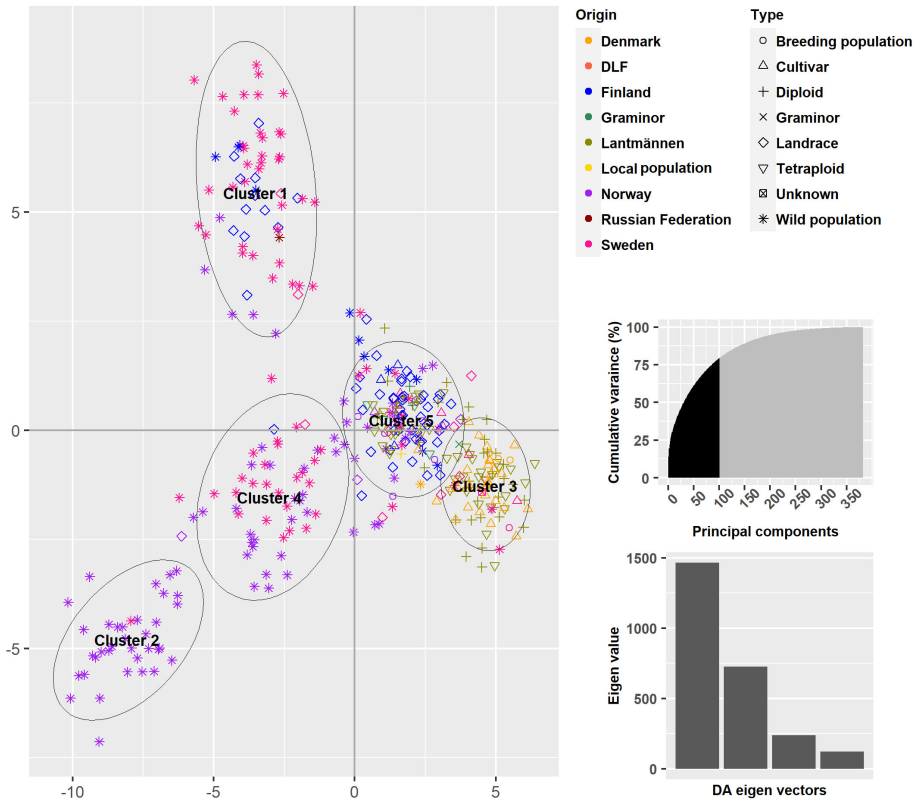


FIGURE 5
A Discriminant Analysis of Principal comonents using 150 Principal components and a five cluster solution.

in the results of LASSO model-based analysis where genetic variation was linked to bioclimatic variables and to relevant biological functions.

Genetic Parameters: H_S, F_{ST}, Nei's standard genetic distance and Tajima's D

The wild populations selected for this study fully spanned the Nordic Region of Europe (Figure 7) with no obvious geographical groupings. The higher F_{ST} (mean of 0.04, Table 1 and Figure 1B) values of the wild populations, compared to the cultivated types, suggests population structure as a consequence of either restricted gene flow, ongoing evolution, or both. In contrast, the cultivated red clover showed low F_{ST} values both within and between different cultivated types and the origins where these groups dominated (Table 1 and Figures 1B, 2A). This indicates a high gene flow within the different types of both the same and different origins. If the F_{ST} between a pair of populations is high, it implies significant differentiation between them. This means that their genetic constitution is significantly different, and hence their crossbreeding may lead to hybrids that are superior to both of them. Since low values of F_{ST} was recorded within cultivated types, crossbreeding with wild populations could lead to further genetic gain.

A lack of genetic differentiation between populations might lead to little to no genetic gain when crossbreeding. This is because the populations possess the same alleles in similar proportions across a majority of loci, and hence crossbreeding does not lead to significant genetic recombination. Even though red clover populations are expected to be highly heterozygous due to their outcrossing nature, variation between populations declines as the majority of their common alleles approach fixation. Tajima's D can be used to measure the amount of rare alleles in a population. Hence, maintaining genetic gain in breeding populations is dependent on the inclusion of new rare alleles. Populations with negative Tajima's D values can be considered to be in expansion following either a bottleneck or selective sweep and thus has an abundance of rare alleles (Tajima, 1989). By selecting such populations further genetic gain can be introduced into the breeding populations. The Tajima's D for the wild populations was negative, which supports the suggestion that an ongoing evolutionary process contributes to their higher genetic differentiation. Hence, the red clover wild populations might have experienced recent selective sweeps and/or events that reduced their population sizes.

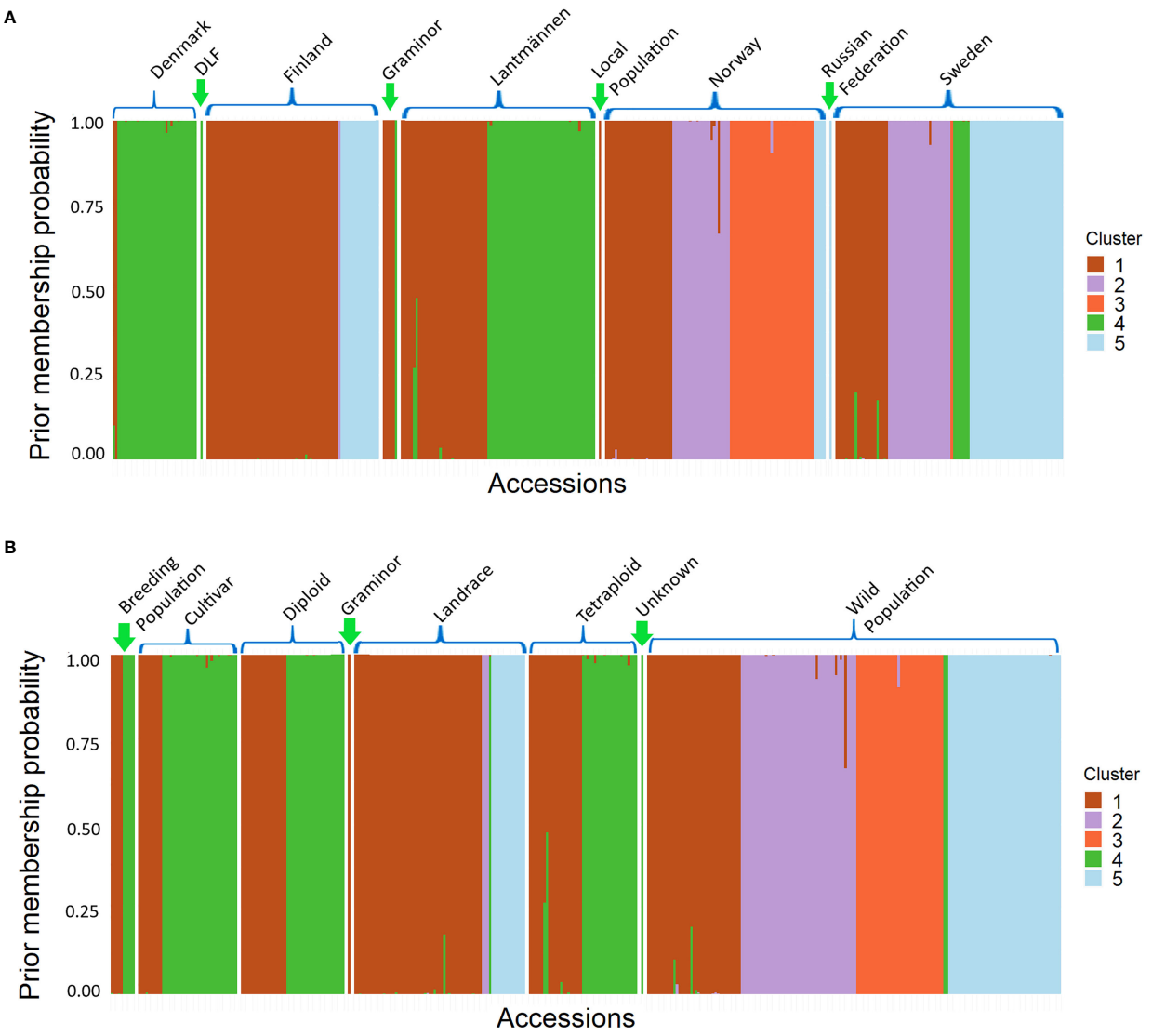


FIGURE 6
A Structure-like membership probability graph showing individual populations of different groups, classified according to **(A)** their origin and **(B)** their population type, assigned to a cluster.

Interestingly, the cultivars and breeding populations from NordGen had lower Tajima's D mean values than the wild populations (Table 1). The lowest Tajima's D values belonged to the Graminor accessions (both as origin and population type, Table 1). This might be due to balancing selection after the development of new cultivars or cultivar types.

Compared to the mean F_{ST} value presented by Jones et al. (2020) of 0.076 for red clover representing Europe and Asia, the mean F_{ST} of the present study (0.022) indicates lower genetic differentiation in Northern European red clover. These results are not surprising since Europe and Asia cover a larger area. Furthermore, red clover was introduced to northern Europe relatively late compared to southern and central Europe (Taylor and Quesenberry, 1996a). Additionally, Jones et al. (2020) used 93.3% ecotypes (here referred to as wild populations) compared to the 45% used in the present study. The results showed that the majority of the genetic diversity was held within the wild populations. Hence, a larger amount of wild populations would increase the F_{ST} in a sample set.

The lowest H_S median was recorded in the wild populations, which also had a higher genetic distance between populations compared to the other population types. The lower Hs values of the wild populations compared to the cultivated types indicate a low gene flow. Of the 382 populations selected for this study, 172 were wild populations. The large proportion of wild populations could be the reason for the relatively lower mean H_S of 0.18 in

the present study (Supplementary Table 1) as compared to Osterman et al. (2021) who reported a mean H_S value of 0.21. Interestingly Jones et al. (2020) reported a mean H_S of 0.26. Therefore, the present study may suggest lower within-population diversity in Norther European wild red clover.

Principal coordinate analysis, discriminant analysis of principal components and neighbor-joining cluster analysis

The PCoA scatter plot for the 382 populations showed a separation between cultivated and wild populations on the first principal coordinate (Figure 3A). The second principal coordinate differentiated Swedish and Finnish wild populations from Norwegian wild populations. The landraces from Sweden and Norway clustered together with NordGen and Lantmännen cultivars while Finnish landrace populations were closer to the Swedish and Finnish wild populations. This pattern was also observed in the DAPC, where wild populations and landraces that were not in clusters 1 and 4 exhibited three major trends, Swedish and Finnish, Swedish and Norwegian, and only Norwegian (Figure 5 and Supplementary Table S1). These two main patterns were observed throughout the analysis, the separation of the Swedish wild populations from the Norwegian wild populations and the mixture of the NordGen and Lantmännen cultivars. This pattern was clearly observed in

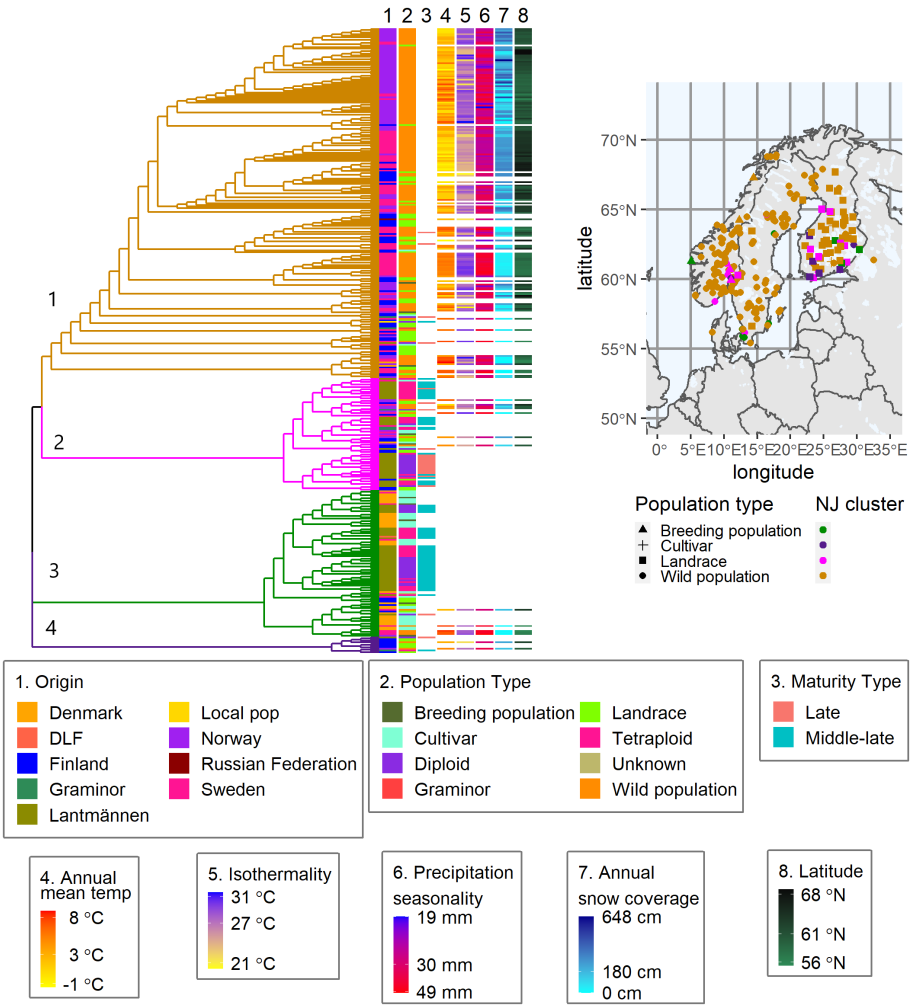


FIGURE 7
A Nei's standard genetic distance based neighbor-joining tree of the 382 populations showing four major clusters. Each column number (1-8) links the column to a specific trait or descriptive data, 1-2 origin and type, 3 maturity types on previously scored populations. Column 4-8 is the bioclimatic variables describing the collection sites of each wild population. The geographical map shows the collection site coordinates for breeding populations, cultivars, landraces, and wild populations obtained from NordGen.

TABLE 2 A summary of the results of best preforming least absolute shrinkage and selection operator (LASSO) model and gene ontology (GO) functional enrichment analysis for the most significant single nucleotide polymorphisms (SNPs).

Environmental parameter	SD	LAMBDA	MAE	RMSE (L)	RMSE (LM)	GO	#	Go genes
Annual mean temperature	2.4	0.2	1.5	1.8	2.4	MF: kinase binding	40	8.3%
Annual precipitation	226	3.7	127	175	227	BP: stomatal opening	88	4.2%
Isothermality	2.3	0.1	1.6	2.2	2.3	None	64	
Latitude	3.1	0.2	1.7	2.1	3	None	63	
Precipitation seasonality	6.1	0.4	4.1	5.3	6.1	None	41	
Annual snow coverage	132	2.5	90	133	131	None	45	
Temperature annual range	4.4	0.1	2.3	3.0	4.4	MF: protein binding BP: stomatal opening	74	50.8%, 4.9%

SD, standard deviation of the input data for each set of bioclimatic variables; RMSE (L), root mean square error of the LASSO model; RMSE (LM), root mean square error of a linear regression; MAE, mean absolute error of the model; GO, gene ontology; #, total number of genes; GO genes, percentage of genes that showed enrichment for each result of LASSO model.

the NJ analysis (Figure 7 and Supplementary Table S1) as well as in Nei's heatmap (Figure 2). This distinction between cultivated and wild red clover suggests that wild populations possess genetic variation that is not represented in the cultivated populations. Hence, by incorporating wild (Norwegian or Swedish) and landrace (Finnish) populations into the breeding programs for red clover, the genetic diversity of the cultivated gene pool can be increased further.

In agreement with the results in Osterman et al. (2021), higher genetic variation differentiated wild populations from Sweden and Norway than the genetic variation that differentiated NordGen and Lantmännen cultivars or diploids and tetraploids. In order to determine the genetic relationship between populations within different groups, various analyses were conducted by grouping the 382 populations according to their origins or types, to reveal any sub-groupings. A separate PCoA analysis for the Lantmännen populations showed no clear differentiation between diploids and tetraploids, in contrast to the low degree of differentiation observed among the NordGen cultivars and landrace populations (Figures 3C, D). Additionally, the lack of clear differentiation between the diploids and tetraploids was evident in the DAPC, where the diploids and tetraploids were assigned to clusters 1 and 2 similarly. However, genetic variation was higher within the diploid group than within the tetraploid group. This can be seen from the PCoA scatter plot where tetraploids were distributed close to the origin while diploids covered the full range of variance described by both PCo1 and PCo2.

In agreement with the results of the PCoA and DAPC, no clear differentiation between diploids and tetraploids was described by Nei's standard genetic distance (Figure 2). However, some sub-clustering of diploids and tetraploids was observed in cluster-2 and cluster-3 of the NJ tree although it was not as clear as their clustering pattern observed in Osterman et al. (2021). The grouping of the tetraploids into different sub-clusters in the present study indicates that they have been derived from chromosome-doubling experiments performed independently on diploids from different genetic backgrounds. Thus, crossbreeding of tetraploids representing different sub-clusters may result in superior cultivars with multiple desirable characteristics. Tetraploid cultivars have higher resilience and biomass yields than diploid cultivars. Hence, genetic differentiation between diploids and tetraploids is expected although it was not the case in the present study. It is likely that the agricultural gain from cultivating tetraploids derives from the molecular genetics of polyploidy rather than from an increased genetic variation since there is no clear genetic variation separating cultivars based on ploidy.

In the case of NordGen germplasm, it is interesting to note that genetic variation was higher among landraces than among cultivars, as clearly depicted in Figure 2D. The two groups also showed significant genetic differentiation, particularly when comparing the landraces from Finland and the cultivars from Denmark (Figures 2D, 5). The distinctness of some Finnish landrace populations was also demonstrated in the heatmap of Nei's genetic distance (Figure 3D) as well as in cluster-4 in the NJ tree (Figure 7). Hence, these Finnish landrace populations might have unique genetic constitution of significant breeding values (agronomic and forage quality) that needs to be explored further. Another interesting finding of the present study was the close genetic relationship between the Danish cultivars from NordGen and the Lantmännen populations (Figures 2, 4). Possibly, this is due to the frequent inclusion of Danish cultivars in Lantmännen breeding programs or to the use of similar genetic resources by different breeding programs to develop cultivars that share similar desirable traits, such as high forage yield.

In order to understand the genetic merit of the gene pool of wild populations, the bioclimatic variables of their respective collection sites were analyzed as a means of examining their respective environments. Wild populations, even naturalized cultivars, are thought to be well adapted to the climate of their natural habitats (Turesson, 1925). Hence, a high genetic

similarity between a cultivar and a wild population may indicate that the cultivar is well suited to an environment similar to that of the wild population. Such analyses can provide insight into whether the germplasm under cultivation has sufficient genetic diversity to suit the diverse environments in which they are being cultivated. The present study revealed that cultivated red clover showed a greater tendency to cluster together with wild populations found in the warmer climates of the south and central parts of the Nordic Region with low levels of variation in precipitation and temperature and little to no snow cover. One of the main causes of red clover senescence is repeated freezing and thawing (Smith, 1957; Zanotto et al., 2021). However, such information is difficult to model. Instead, an educated guess can be made using isothermality and snow coverage to identify locations that may have long autumns with frequent fluctuations around the freezing point. Wild populations from northern Norway in cluster 2 (Figure 7), where the snow coverage is high and the annual mean temperature is around zero or negative, shared a close genetic relationship with three middle-late diploid cultivars bred by Lantmännen. Hence, it would be interesting to evaluate the winter hardiness of these cultivars to validate the ideas discussed above.

LASSO prediction and GO functional analysis

This study used least absolute shrinkage and selection operator (LASSO) models to relate the SNP frequency across populations to a specific bioclimatic variable. Due to its ability to rank the importance of variables LASSO models are currently used in multiple fields where the number of samples is less than the number of variables. They are used in gene-based diagnostics (Kohannim et al., 2012; Kim et al., 2018), genome-wide association research (Li et al., 2011), and other forms of unsupervised learning like in chemometrics (Pomareda et al., 2010). In the present study, the objective was to identify highly descriptive markers that can help select suitable germplasm for use in breeding programs. To the best of our knowledge, the LASSO models have not been used to relate a SNP marker to an environmental variable before. In this study, the method was considered successful because it was possible to assess the relationship between the SNPs identified by the LASSO models and the traits appropriate to the bioclimatic variable studied. Hence, this study demonstrates the ability of penalized linear regression models to assess the relationship between SNPs and environments.

Relating SNPs to bioclimatic variables with allele frequencies have previously been done via Bayesian models. However, for these models to converge, the data must fulfill the assumptions of the prior likelihood distribution. If the data does not fit the Bayesian model, a LASSO model could be used as an alternative as they rely on no prior information. LASSO models, however, are not adjusted based on population structure, unlike Bayesian models. As a result, additional steps are necessary in order to exclude possible artifacts due to population structure or statistical false discovery. A GO enrichment analysis was carried out for this purpose in the present study and satisfactory results were obtained.

Four LASSO models showed enrichment for biological processes that can be regarded as plausible responses of plants that are growing in a particular environment (Table 2). For example, the temperature and precipitation models showed enrichment of genes related to stomatal opening, a function known to be involved in the plant response to humidity and temperature (Waggoner and Zelitch, 1965; Honour et al., 1995). Additionally, the enrichment of protein kinases in the annual mean temperature model gives further information on the mechanisms of plant resilience. For a better understanding of persistence in red clover, a study of the stomatal changes and kinase activities in plants bearing different alleles related to their survival would be valuable. Similarly, the knowledge of how wild red clover copes with temperature stress is valuable for breeding since cold resilience is a desirable trait in Nordic breeding programs and beyond. Additionally, a study on root development between cultivated and wild red clover would be of interest given the results of the GO enrichment of auxiliary root development. The establishment of roots could affect persistence, resilience, nitrogen fixation, and nutrient uptake, as well as the establishment of the whole plant. This way of evaluating new germplasm could serve as a key component of red clover improvement.

Conclusion

This study thoroughly described the genetic diversity and population structure of the Nordic red clover genetic resources, which include breeding populations, cultivars, landraces, and wild populations. As shown by this study, further genetic gains are possible by incorporating NordGen cultivars and landraces. Inclusion of selected landraces and wild populations based on the results exhibited in Figure 7 into red clover breeding programs could increase persistence and climate resilience of cultivars and synthetic populations. Furthermore, GO enrichment analysis facilitated the identification of SNPs that may affect the stomatal function and root development in wild populations, thus providing additional knowledge for breeding this forage crop. It would be very interesting to see this method applied in other similar studies involving wild germplasm.

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/; PRJNA765476.

Author contributions

MG secured the funding with help from RO, CH, and other project participants. With inputs from CH and RO, MG designed the genotyping part of the study while JO planned the statistical analyses regarding bioclimatic variables. JO, CH, and MG conducted the greenhouse work, including sampling leaf tissue for DNA extraction. JO wrote the manuscript draft and revised it. MG, CH, and RO reviewed the different versions of the draft and assisted in the revision process. All authors contributed to the article and approved the submitted version.

Funding

The study was fully funded by SLU Grogrund – Centre for Breeding of Food Crops, Swedish University of Agricultural Sciences.

Acknowledgments

We would like to thank Linda Öhlund (Lantmännen) for providing us with seeds of red clover cultivars and breeding populations. We would also like to thank NordGen for supplying us with selected red clover germplasm from their collection. We are thankful to Mohammad El-Khalifeh (NordGen) for the additional help in selecting the accessions, and to the greenhouse management staff for keeping the plants healthy. We would like to thank David Parsons (SLU), Elisabet Nadeau (SLU) and Alf Ceplitis (Lantmännen) for constructive comments on the results of the study during the project meetings.

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.2022.997860/full#supplementary-material

References

Amdahl, H., Aamlid, T. S., Ergon, Å., Kovi, M. R., Marum, P., Alsheikh, M., et al. (2016). Seed yield of Norwegian and Swedish tetraploid red clover (Trifolium pratense l.) populations. *Crop Sci.* 56, 603–612. doi: 10.2135/cropsci2015.07.0441

Andri, S. (2021) *DescTools: Tools for descriptive statistics*. Available at: https://cran.r-project.org/package=DescTools (Accessed 1st January 2022).

Benevenuto, J., Bhakta, M., Lohr, D. A., Ferrão, L. F. V., Resende, M. F. R., Kirst, M., et al. (2019). Cost-effective detection of genome-wide signatures for 2,4-d herbicide resistance adaptation in red clover. *Sci. Rep.* 9, 20037. doi: 10.1038/s41598-019-55676-9

Colby, T., Matthäi, A., Boeckelmann, A., and Stuible, H.-P. (2006). SUMO-conjugating and SUMO-deconjugating enzymes from arabidopsis. *Plant Physiol.* 142, 318–332. doi: 10.1104/pp.106.085415

De Vega, J. J., Ayling, S., Hegarty, M., Kudrna, D., Goicoechea, J. L., Ergon, Å., et al. (2015). Red clover (trifolium pratense l.) draft genome provides a platform for trait improvement. *Sci. Rep.* 5, 17394. doi: 10.1038/srep17394

Dhamala, N. R., Eriksen, J., Carlsson, G., Søegaard, K., and Rasmussen, J. (2017). Highly productive forage legume stands show no positive biodiversity effect on yield and N2-fixation. *Plant Soil* 417, 169–182. doi: 10.1007/s11104-017-3249-2

Ergon, Å., Skøt, L., Sæther, V. E., and Rognli, O. A. (2019). Allele frequency changes provide evidence for selection and identification of candidate loci for survival in red clover (Trifolium pratense l.). *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00718

Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatology* 37, 4302–4315. doi: 10.1002/joc.5086

Gautier, M. (2015). Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201, 1555–1579. doi: 10.1534/genetics.115.181453

Gautier, M., Vitalis, R., Flori, L., and Estoup, A. (2022). f-statistics estimation and admixture graph construction with Pool-seq or allele count data using the R package poolfstat. *Molecular Ecology Resources* 32, 1394–1416. doi: 10.1111/1755-0998.13557

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847–2849. doi: 10.1093/bioinformatics/btw313

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512

Herrmann, D., Boller, B., Studer, B., Widmer, F., and Kölliker, R. (2008). Improving persistence in red clover: Insights from QTL analysis and comparative phenotypic evaluation. *Crop Sci.* 48, 269–277. doi: 10.2135/cropsci2007.03.0143

Hijamns, R. J., Etten, J., Sumner, M., Cheng, J., Baston, D., Bevan, A., et al. (2012) *Raster: Geographic data analysis and modeling*. Available at: https://CRAN.R-project.org/package=raster (Accessed December 3, 2021).

Hirsch, C. N., Hirsch, C. D., Felcher, K., Coombs, J., Zarka, D., Van Deynze, A., et al. (2013). Retrospective view of north American potato (Solanum tuberosum l.) breeding in the 20th and 21st centuries. *G3 Genes|Genomes|Genetics* 3, 1003–1013. doi: 10.1534/g3.113.005595

Honour, S. J., Webb, A. A. R., and Mansfield, T. A. (1995). The responses of stomata to abscisic acid and temperature are interrelated. *Proc. R. Soc. London. Ser. B: Biol. Sci.* 259, 301–306. doi: 10.1098/rspb.1995.0044

Jombart, T. (2008). Adegenet: A r package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Collins, C. (2015). A tutorial for discriminant analysis of principal components (DAPC) using adegenet.

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-11-94

Jones, C., De Vega, J., Lloyd, D., Hegarty, M., Ayling, S., Powell, W., et al. (2020). Population structure and genetic diversity in red clover (trifolium pratense l.) germplasm. *Sci. Rep.* 10, 8364. doi: 10.1038/s41598-020-64989-z

Kim, S. M., Kim, Y., Jeong, K., Jeong, H., and Kim, J. (2018). Logistic LASSO regression for the diagnosis of breast cancer using clinical demographic data and the BI-RADS lexicon for ultrasonography. *Ultrasonography* 37, 36–42. doi: 10.14366/usg.16045

Kofler, R., Orozco-terWengel, P., Maio, N. D., Pandey, R. V., Nolte, V., Futschik, A., et al. (2011a). PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PloS One* 6, e15925. doi: 10.1371/journal.pone.0015925

Kofler, R., Pandey, R. V., and Schlötterer, C. (2011b). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-seq). *Bioinformatics* 27, 3435–3436. doi: 10.1093/bioinformatics/btr589

Kohannim, O., Hibar, D., Stein, J., Jahanshad, N., Hua, X., Rajagopalan, P., et al. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Front. Neurosci.* 6. doi: 10.3389/fnins.2012.00115

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28, 1–26. doi: 10.18637/jss.v028.i05.

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756. doi: 10.1093/genetics/124.3.743

Larkindale, J., Hall, J. D., Knight, M. R., and Vierling, E. (2005). Heat stress phenotypes of arabidopsis mutants implicate multiple signaling pathways in the acquisition of thermotolerance. *Plant Physiol.* 138, 882–897. doi: 10.1104/pp.105.062257

Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* 27, 516–523. doi: 10.1093/bioinformatics/btq688

Li, W., Riday, H., Riehle, C., Edwards, A., and Dinkins, R. (2019). Identification of single nucleotide polymorphism in red clover (Trifolium pratense l.) using targeted genomic amplicon sequencing and RNA-seq. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01257

Luu, K., Bazin, E., and Blum, M. G. B. (2017). Pcadapt: An r package to perform genome scans for selection based on principal component analysis. *Mol. Ecol. Resour.* 17, 67–77. doi: 10.1111/1755-0998.12592

McKenna, P., Cannon, N., Conway, J., and Dooley, J. (2018). The use of red clover (Trifolium pratense) in soil fertility-building: A review. *Field Crops Res.* 221, 38–49. doi: 10.1016/j.fcr.2018.02.006

Öhberg, H. (2008) *Studies of the persistence of red clover cultivars in Sweden*. Available at: https://pub.epsilon.slu.se/1741/ (Accessed July 7, 2021).

Osterman, J., Hammenhag, C., Ortiz, R., and Geleta, M. (2021). Insights into the genetic diversity of Nordic red clover (Trifolium pratense) revealed by SeqSNP-based genic markers. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.748750

Pandey, J., Scheuring, D. C., Koym, J. W., Coombs, J., Novy, R. G., Thompson, A. L., et al. (2021). Genetic diversity and population structure of advanced clones selected over forty years by a potato breeding program in the USA. *Sci. Rep.* 11, 8344. doi: 10.1038/s41598-021-87284-x

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in r language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Poczai, P., Varga, I., Laos, M., Cseh, A., Bell, N., Valkonen, J. P., et al. (2013). Advances in plant gene-targeted and functional markers: A review. *Plant Methods* 9, 6. doi: 10.1186/1746-4811-9-6

Pomareda, V., Calvo, D., Pardo, A., and Marco, S. (2010). Hard modeling multivariate curve resolution using LASSO: Application to ion mobility spectra. *Chemometrics Intelligent Lab. Syst.* 104, 318–332. doi: 10.1016/j.chemolab.2010.09.010

Praat, M., De Smet, I., and van Zanten, M. (2021). Protein kinase and phosphatase control of plant temperature responses. *J. Exp. Bot.* 72, 7459–7473. doi: 10.1093/jxb/erab345

R Core Team (2013). *R: A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing). Available at: http://www.R-project.org/.

Schlötterer, C., Tobler, R., Kofler, R., and Nolte, V. (2014). Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* 15, 749–763. doi: 10.1038/nrg3803

Selga, C., Chrominski, P., Carlson-Nilsson, U., Andersson, M., Chawade, A., and Ortiz, R. (2022). Diversity and population structure of Nordic potato cultivars and breeding clones. *BMC Plant Biology* 22, 350. doi: 10.1186/s12870-022-03726-2

Smith, D. (1957). Flowering response and winter survival in seedling stands of medium red Clover1. *Agron. J.* 49, 126–129. doi: 10.2134/agronj1957.00021962004900030005x

Smith, R. R., Taylor, N. L., and Bowley, S. R. (1985). "Red clover," in *Clover science and technology* (John Wiley & Sons, Ltd), 457–470. doi: 10.2134/agronmonogr25.c19

Sturz, A. V., Christie, B. R., Matheson, B. G., and Nowak, J. (1997). Biodiversity of endophytic bacteria which colonize red clover nodules, roots, stems and foliage and their influence on host growth. *Biol. Fertil Soils* 25, 13–19. doi: 10.1007/s003740050273

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Taylor, N. L., and Quesenberry, K. H. (1996a). "Historical perspectives," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 1–10. doi: 10.1007/978-94-015-8692-4_1

Taylor, N. L., and Quesenberry, K. H. (1996b). "Persistence," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 119–129. doi: 10.1007/978-94-015-8692-4_10

Taylor, N. L., and Quesenberry, K. H. (1996c). "Tetraploid red clover," in *Red clover science current plant science and biotechnology in agriculture*. Eds. N. L. Taylor and K. H. Quesenberry (Dordrecht: Springer Netherlands), 161–169. doi: 10.1007/978-94-015-8692-4_13

Thilakarathna, M. S., Papadopoulos, Y. A., Grimmett, M., Fillmore, S. A. E., Crouse, M., and Prithiviraj, B. (2017). Red clover varieties show nitrogen fixing advantage during the early stages of seedling development. *Can. J. Plant Science.* 98, 517–526. doi: 10.1139/cjps-2017-0071

Turesson, G. (1925). The plant species in relation to habitat and climate. *Hereditas* 6, 147–236. doi: 10.1111/j.1601-5223.1925.tb03139.x

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2021). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res.* 50, D1468–1474. doi: 10.1093/nar/gkab1024

Vitalis, R., Gautier, M., Dawson, K. J., and Beaumont, M. A. (2014). Detecting and measuring selection from gene frequency data. *Genetics* 196, 799–817. doi: 10.1534/genetics.113.152991

Waggoner, P. E., and Zelitch, I. (1965). Transpiration and the stomata of leaves. *Science* 150, 1413–1420. doi: 10.1126/science.150.3702.1413

Wu, J., Wang, J., Pan, C., Guan, X., Wang, Y., Liu, S., et al. (2014). Genome-wide identification of MAPKK and MAPKKK gene families in tomato and transcriptional profiling analysis during development and stress response. *PloS One* 9, e103032. doi: 10.1371/journal.pone.0103032

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2017). Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zanotto, S., Palmé, A., Helgadóttir, Á., Daugstad, K., Isolahti, M., Öhlund, L., et al. (2021). Trait characterization of genetic resources reveals useful variation for the improvement of cultivated Nordic red clover. *J. Agron. Crop Sci.* 207, 492–503. doi: 10.1111/jac.12487

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 13 October 2022
DOI 10.3389/fpls.2022.1015658

OPEN ACCESS

EDITED BY
Andrés J. Cortés,
Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

REVIEWED BY
Ofere Francis Emeriewen,
Julius Kühn Institute (JKI), Germany
Satish Kumar,
The New Zealand Institute for Plant and Food Research Ltd, New Zealand

*CORRESPONDENCE
Gayle M. Volk
Gayle.Volk@usda.gov

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

RECEIVED 09 August 2022
ACCEPTED 21 September 2022
PUBLISHED 13 October 2022

CITATION
Volk GM, Peace CP, Henk AD and Howard NP (2022) DNA profiling with the 20K apple SNP array reveals *Malus domestica* hybridization and admixture in *M. sieversii*, *M. orientalis*, and *M. sylvestris* genebank accessions. *Front. Plant Sci.* 13:1015658. doi: 10.3389/fpls.2022.1015658

DNA profiling with the 20K apple SNP array reveals *Malus domestica* hybridization and admixture in *M. sieversii*, *M. orientalis*, and *M. sylvestris* genebank accessions

Gayle M. Volk[1]*, Cameron P. Peace[2], Adam D. Henk[1] and Nicholas P. Howard[3]

[1]United States Department of Agriculture-Agricultural Research Service (USDA-ARS) National Laboratory for Genetic Resources Preservation, Fort Collins, CO, United States, [2]Department of Horticulture, Washington State University, Pullman, WA, United States, [3]Fresh Forward Breeding and Marketing B.V., Huissen, Netherlands

The USDA-ARS National Plant Germplasm System (NPGS) apple collection in Geneva, NY, USA maintains accessions of the primary *Malus domestica* (Suckow) Borkh. progenitor species *M. sieversii* (Ledeb.) M. Roem., *M. orientalis* Uglitzk., and *M. sylvestris* (L.) Mill. Many of these accessions originated from seeds that were collected from wild populations in the species' centers of diversity. Some of these accessions have fruit phenotypes that suggest recent *M. domestica* hybridization, which if true would represent crop contamination of wild species populations and mislabeled species status of NPGS accessions. Pedigree connections and admixture between *M. domestica* and its progenitor species can be readily identified with apple SNP array data, despite such arrays not being designed for these purposes. To investigate species purity, most (463 accessions) of the NPGS accessions labeled as these three progenitor species were genotyped using the 20K apple SNP array. DNA profiles obtained were compared with a dataset of more than 5000 unique *M. domestica* apple cultivars. Only 212 accessions (151 *M. sieversii*, 26 *M. orientalis*, and 35 *M. sylvestris*) were identified as "pure" species representatives because their DNA profiles did not exhibit genotypic signatures of recent hybridization with *M. domestica*. Twenty-one accessions (17 *M. sieversii*, 1 *M. orientalis*, and 3 *M. sylvestris*) previously labeled as wild species were instead fully *M. domestica*. Previously unrealized hybridization and admixture between wild species and *M. domestica* was identified in 230 accessions (215 *M. sieversii*, 9 *M. orientalis*, and 6 *M. sylvestris*). Among these species-mislabeled accessions, 'Alexander', 'Gold Reinette', 'Charlamoff', 'Rosmarina Bianca', and 'King of the Pippins' were the most frequently detected *M. domestica* parents or grandparents. These results have implications for collection management, including germplasm distribution,

and might affect conclusions of previous research focused on these three progenitor species in the NPGS apple collection. Specifically, accessions received from the NPGS for breeding and genomics, genetics, and evolutionary biology research might not be truly representative of their previously assigned species.

KEYWORDS

crop wild relatives, cultivar, genetic diversity, genotype, Central Asia

1 Introduction

Apple (*Malus*) genebank collections make plant genetic resources available to research and breeding programs that seek new alleles for improving disease and pest resistance, reducing environmental vulnerabilities, and improving production and consumer traits (Volk et al., 2015a; Bramel and Volk, 2019; Peace et al., 2019). Genebanks include *Malus domestica* cultivars as well as accessions that represent diverse species. For breeding, closely related *Malus* crop wild relatives provide desirable alleles without the extreme challenges of working with more distant wild species (Migicovsky and Myles, 2017). Researchers and breeders depend on apple collections to provide high quality materials that are true-to-type, at both the species and cultivar levels.

Donations, exchanges, and plant explorations have made the USDA National Plant Germplasm System (NPGS) apple collection among the largest and most diverse collections in the world (Volk et al., 2015a; Bramel and Volk, 2019; Gutierrez et al., 2020). The collection provides a wide range of *Malus* crop wild relatives that have been used to determine evolutionary relationships, identify novel alleles, and assess genetic diversity (Volk et al., 2015a). These crop wild relatives, including the primary *M. domestica* progenitor species of *M. sieversii*, *M. sylvestris*, and *M. orientalis*, were acquired by the NPGS through plant exchange and exploration expeditions that were performed between 1989 and 2004 (Luby et al., 2001; Forsline et al, 2002; Volk et al., 2009b). Exploration efforts introduced either budwood or seeds into the NPGS after passing through the United States national quarantine program. Budwood was grafted onto rootstocks and some seedlots were planted to obtain seedling trees in orchard blocks in Geneva, NY. Many of the trees have been genotyped using a set of 7 or 19 microsatellite markers, and the results were used to identify core subsets and genetic relationships among accessions (Richards et al., 2009a; Richards et al., 2009b; Volk et al., 2005; Volk et al., 2009b). Core subset and elite accessions (those exhibiting unusual or desirable phenotypes) were propagated by grafting and included in the permanent orchard collection. In addition, seedling trees produced from crosses between 'Royal Gala' and *M. sieversii*-labeled accessions PI 613971 (GMAL 4327), PI 613981 (GMAL 4448), and PI 613988 (GMAL 4455) resulted in research populations with local identifiers GMAL 4590, GMAL 4593, and GMAL 4595, respectively. These populations have been used for multiple genetic linkage mapping studies involving the *Ma* locus influencing fruit acidity (Xu et al., 2012), resistance to blue mold (Norelli et al., 2014; Norelli et al., 2017), and resistance to apple scab (Wang et al., 2012).

Malus sieversii, native to Central Asia and Western China, offers novel allelic diversity for a plethora of traits that are important to breeding programs (Volk et al., 2015a; Liu et al., 2021). A primary progenitor species of *M. domestica*, *M. sieversii* has been the target of numerous studies that have focused either on materials from China or Central Asia (Zhang et al., 2007). The extensive NPGS collection of this species in Geneva, NY, was obtained by Phil Forsline, Herb Aldwinckle, A. D. Dzhangaliev, and their collaborators in Kazakhstan and other Central Asian countries in 1989, 1993, 1995, and 1996. Four collection trips resulted in a total of 894 seedlots, and hundreds of these seeds were planted in the Geneva orchards, with many others provided to collaborators both within the U.S. and abroad (Forsline et al., 2002). Some trees in the wild in Kazakhstan had fruit quality phenotypes that rivaled those of cultivars and were therefore considered "elites" during the 1990s collection trips. Johann Sievers, after whom *M. sieversii* was named, during his travels in 1790 described wild apples of Kazakhstan (near Ust-Kamenogorsk, 500 km northeast of "Site 9" and about 1100 km northeast of Almaty) as dwarf trees with apples the size of a chicken egg, having red and yellow cheeks, and that could be eaten from the trees (Nussenov, 2018). This suggests that, as early as 1790, larger wild *M. sieversii* apples were present in the Targabatai Mountains and other northeast regions of Kazakhstan. "Elites" and other trees were introduced into the NPGS as budwood and grafted onto rootstocks, while most *M. sieversii* accessions were introduced as seed. *M. sieversii* germplasm accessioned into the NPGS has been distributed as budwood of accessions resulting from imported budwood of

wild trees considered "elites", budwood of accessions grown from wild-collected seeds, and seed from crossing among wild accessions grown ex situ. These materials have subsequently been used to determine genetic relationships between *M. domestica* cultivars and *M. sieversii* (Robinson et al., 2001; Gharghani et al., 2009; Nikiforova et al., 2013; Duan et al., 2017; Wedger et al., 2021), to evaluate phenotypic diversity of traits (Janisiewicz et al., 2008; Fazio et al., 2009; Bassett et al., 2011; Jurick et al., 2011; Van Nocker et al., 2012; Fazio et al., 2014; Maguylo and Bassett, 2014; Harshman et al., 2017; Watts et al., 2021; Davies et al., 2022), and to identify QTLs and novel alleles (Xu et al., 2012; Wisniewski et al., 2020; Singh et al., 2021).

Malus orientalis, native to the Caucasus and Middle East, is also a likely contributor to the domesticated apple (Cornille et al., 2012; Cornille et al., 2014; Amirchakhmahgi et al., 2018). Its attributes of interest to breeding programs include late blooming, environmental adaptation, fire blight resistance (Amirchakhmaghi et al., 2022), and long-term storage (Khadivi et al., 2020; Moradi et al., 2022). *M. orientalis* in the NPGS was primarily collected in exploration trips to Turkey, Russia, Armenia, and the Republic of Georgia between 1998 and 2004 (Volk et al., 2009b). The NPGS accessions of *M. orientalis* have been used for genetic and phenotypic research (Gharghani et al., 2009; Volk et al., 2009b; Duan et al., 2017).

Malus sylvestris, found in localized wild populations throughout much of Europe, has increasingly become recognized as an important ancestral contributor to *M. domestica*, but has not been as extensively utilized in breeding and research as the other two main progenitor species (Cornille et al, 2012; Cornille et al., 2013; Duan et al., 2017). Much of the literature on *M. sylvestris* has instead focused on the issue of recent hybridization with *M. domestica* cultivars, which is reportedly rife in wild populations. Such hybridization has been identified in the East Ore Mountains of Germany (40% of trees being hybrids; Reim et al., 2013), Saxony (13% hybrids; Reim et al., 2020), the Rhine Valley (5% hybrids; Schnitzler et al., 2014), and the United Kingdom (30% hybrids or pure *M. domestica*; Ruhsam et al., 2019). A recent study identified seven of 115 *M. sylvestris* accessions to be admixed with *M. domestica* in The Netherlands' field genebank collection (Buiteveld et al., 2021).

SNP arrays have become a very powerful tool in apple, particularly for use in haplotype-based analyses such as pedigree-based QTL analyses (Kostick and Luby, 2022), introgression tracking (Luo et al., 2020), and relatedness estimation (Howard et al., 2021a). Several SNP arrays have been developed for use in apple, with the Illumina Infinium® 20K (Bianco et al., 2014) being the most common. While these SNP arrays were designed using panels consisting primarily of *M. domestica* cultivars without input from any wild *M. sieversii*, *M. orientalis*, or *M. sylvestris*, it is expected that they will work well with these species being the progenitors of domesticated apple and because the 20K SNP array was successfully used for introgression tracking of *M. sieversii* haplotypes (Luo et al., 2020).

Acknowledging the importance of providing true-to-type species materials in the NPGS apple collection, the purpose of this study was to provide accurate information about the extent of hybridization (clear single recent crossing events) and admixture (multi-generational species mixing without specific ancestors identified) in NPGS accessions of *M. sieversii*, *M. orientalis*, and *M. sylvestris*, using the Illumina Infinium 20K SNP array.

2 Materials and methods

2.1 Plant material

Leaves were sampled from a total of 383 *M. sieversii*, 36 *M. orientalis*, 44 *M. sylvestris*, and one *M. domestica*-labeled accessions from the USDA National Plant Germplasm System apple collection in Geneva, NY (Table S1; USDA, 2022).

2.2 DNA extraction

Fresh frozen (100 mg) or dried (50 mg) apple leaf tissue was pulverized to a fine powder and DNA extracted using a modified CTAB extraction procedure (File S1). DNA quality and quantity were determined using a spectrophotometer/fluorometer.

2.3 Genotypic analysis

Samples were genotyped on the Illumina Infinium® 20K apple SNP array (Bianco et al., 2014). Raw SNP array data were curated according to Vanderzande et al. (2019). The resulting genome-wide SNP profiles for the accessions were added to a dataset of more than 5000 unique genotypic profiles sampled from 56 apple collections previously assembled for an ongoing collaborative apple pedigree reconstruction project (Howard et al., 2018). *Malus* unique genotype (MUNQ) codes used for the organization of duplicate genotypic profiles were provided *via* Denancé et al. (2020).

Admixture was identified *via* a combination of an analysis of Summed Potential Lengths or Shared Haplotypes (SPLoSH) information (Howard et al., 2021a) and Principal Components Analysis (PCA). The commonly used STRUCTURE analysis (Pritchard et al., 2000) was not used because our pilot study identified extensive pedigree structure between many wild accessions and extant domestic cultivars, because of the presence of extensive pedigree structure among *M. sieversii* accessions, because of the small number of *M. sylvestris* accessions available for analysis, because of the extensive pedigree structure inherent in any panel of domestic cultivars

that could be included in a STRUCTURE analysis, and because of issues relating to the SNP inclusion bias on the 20K SNP array. These issues would have severely violated some of the assumptions made in the STRUCTURE model or would have otherwise resulted in unclear or misleading results.

Genetic duplicates and parent-offspring relationships were sought among the progenitor species accessions and the larger dataset of DNA profiles as described in Vanderzande et al. (2019). Close pedigree relationships and grandparent-grandchild relationships were identified using SPLoSH information as described in Howard et al. (2021a) using 20 cM as a threshold. This threshold was chosen to readily enable detection of any recent cultivar ancestors of the species accessions in the dataset. If one parent of an accession was identified, haplotype data deduced for the chromosomal homologs from the unknown second parent were also compared to the dataset to detect any likely recent cultivar ancestors.

Grandparent-grandchild relationships involving species accessions were considered likely present where the SPLoSH values between pairs were 512 cM or higher, representing at least 20% of the entire diploid genome [twice the 1280 cM haploid genetic length; Howard et al. (2021a)]. These thresholds were used instead of the estimated coefficient of relatedness models from Howard et al. (2021a) for three reasons. First, the coefficient of relatedness model estimates from Howard et al. (2021b) were made only using *M. domestica* cultivars that were expected to have a degree of haplotype sharing through multi-generational endogamy (*via* artificial selection). Such shared ancestry would be expected to inflate the estimated average SPLoSH values for grandparent-grandchild and half-sib relationships among *M. domestica* cultivars. Thus, in instances where species accessions had a *M. domestica* grandparent but otherwise appeared to be of a progenitor species origin, the SPLoSH value between them would not have that inflation and instead would more closely approximate the theoretical 25% of genome sharing for this relationship. Second, the 20-cM threshold used could have prevented detection of some real but shorter identical-by-descent haplotypes, leading to a total amount of genome-sharing less than the expected 25%. But a smaller threshold than 20 cM was not used because those could also have resulted in some false or artificially elongated shared haplotypes due to limitations of the SNP coverage (some gaps being present) and informativeness (undetected null alleles possible) of the 20K array. Third, using a genome-sharing threshold slightly less than 25% allowed for expected biological variation in proportion of genome inherited after two meioses. Thus, the minimum threshold of 512 cM used for likely grandparent-grandchild considered in this study was intended to address these points by limiting both the exclusion of real relationships and the inclusion of false relationships.

Accessions were classified as fully *M. domestica* if their entire pedigree consisted of *M. domestica* cultivars. Accessions were classified as hybrid if they had one *M. domestica* parent or at least one likely *M. domestica* grandparent. Accessions were classified as having a *M. domestica* component, and thus admixed, if they had SPLoSH values with any *M. domestica* cultivar of more than 10% of their genome ($0.1 \times 2 \times 1280$ cM = 256 cM) but could not be classified as hybrid or fully *M. domestica*. If accessions had SPLoSH values with numerous *M. domestica* cultivars that were between 7.5% and 10% of their genome, they were noted as such but not classified as having a *M. domestica* component. Progenitor species accessions lacking definitive or clear evidence of *M. domestica* introgression from SPLoSH information were compared to one another through PCA to identify outliers that could also indicate admixed individuals. PCA was conducted using prcomp in R (R Core team, 2022). SPLoSH information using 5 cM as a threshold was used as the input information instead of raw SNP data to diminish effects of unequal SNP informativeness across the material and chromosomes and to account for genetic linkage among SNPs. PCA results were used to confirm or clarify the recorded species of accessions without clear admixture by observing clustering patterns. Outlier accessions positioned outside but between the primary species clusters were noted as being possible hybrids/admixed between those species. Such accessions were also examined for any abnormally large SPLoSH values with *M. domestica* cultivars relative to those found in accessions that did not have outlier PCA positions to gain evidence for the possibility of smaller-scale admixture.

Some accessions were classified as having an "exotic" *Malus* component. Exotic *Malus* in this study refers to *Malus* species with fruit smaller than *M. orientalis*, *M. sieversii*, and *M. sylvestris* accessions and very long stems, such as *M. baccata* (L.) Borkh., *M. floribunda* Siebold ex Van Houtte, *M.* × *micromalus* Makino, and *M. toringo* (Siebold) de Vriese (also referred to as *M. sieboldii* Rehder), and which are all not primary progenitors of *M. domestica*. Accessions were considered as having an exotic *Malus* component if they shared more than 256 cM (i.e., at least 10% of the diploid genome) of SPLoSH using 20 cM as a threshold with a group of 11 phenotypically confirmed exotic *Malus* accessions (Table S2) that lacked significant SPLoSH values with *M. domestica* cultivars.

Passport details of 28 of the 44 *M. sylvestris* accessions were recorded in GRIN-Global (USDA, 2022; and confirmed by genotypic analyses here) as belonging to three separate full-sibling groups derived from four different parents, only two of which were available for genotyping. To prevent this pedigree structure from causing *M. sylvestris* accession outliers to appear in the PCA, SNP profiles for the two ungenotyped parents, 'Oelsen 5' and 'Klipphausen', were imputed using their recorded offspring *via* the method of Howard et al. (2021a) and the parents of these full-sib groups were used for PCA instead of the full-sibs.

2.4 Phenotypic analysis

Available fruit image data were downloaded from GRIN-Global for the 20K SNP array-genotyped NPGS accessions of *M. sieversii*, *M. orientalis*, and *M. sylvestris.* Additional photograph imaging (Nikon D7100, 4000 × 6000 pixels) of multiple fruit was conducted and then uploaded to GRIN-Global for *M. sieversii*, *M. orientalis*, and *M. sylvestris* genotyped accessions that did not previously have associated image data available. From the 234 images, phenotypic measurements were conducted for five fruit of each accession for the traits of fruit diameter, fruit ground color, percentage of fruits with overcolor, percentage of each fruit with red overcolor, and fruit shape (according to Watkins and Smith, 1997; Figure S1). Quantitative data were analyzed by ANOVA and Tukey Mean Separation tests.

3 Results

3.1 Genotypic analysis

The SNP array performed effectively on accessions of all three progenitor species to detect hybridization and admixture, enable pedigree reconstruction using SPLoSH information, and enable DNA profile imputation of two ungenotyped *M. sylvestris* parents. SPLoSH information was able to reliably illuminate clear instances of admixture, often directly through domestic cultivars. As an illustrative example, *M. domestica-M. sieversii* hybrid PI 613979 was identified as an offspring of 'Alexander' (Figure 1A), and phased haplotypes, with recombination evidence, from 'Alexander' clearly accounted for one homolog of each chromosome of PI 613979. In PI 650959, an offspring of PI 613979, remnant haplotypes of its grandparent 'Alexander' can clearly be identified (Figure 1B).

All accessions with a non-*M. domestica* component tended to have higher numbers of null alleles present than did pure *M. domestica* cultivars, although this did not seem to impede admixture detection. The *M. sieversii* accession with the highest level of data curation *via* descendants, PI 613981, had 41 SNPs detected as being homozygous for null alleles. For *M. orientalis*, the highest detected number of homozygous-null SNPs was 21 (in PI 682807), and the highest detected number in a curated *M. sylvestris* accession was 36 (in GMAL 4495.o, having SNP data available for both parents). PCA based on the SNP array data clearly differentiated the three progenitor species (Figure S2).

3.1.1 *Malus sieversii*

Of the 383 accessions originally labeled as *M. sieversii*, 151 were determined to be pure *M. sieversii*, 17 were determined to be fully *M. domestica*, and 215 were hybrids or admixed between taxa. Specifically regarding the latter, 178 accessions were *M. sieversii-M. domestica* hybrids, 33 were *M. sieversii* with domestic components, one was a *M. sieversii-M. orientalis* hybrid, and three were *M. sieversii* with an exotic component (Table S1). In addition, one *M. domestica*-labeled accession (PI 644151) was determined to be a *M. sieversii-M. orientalis* hybrid. In all, only 39% of the sampled accessions labeled as *M. sieversii* in the NPGS genebank were determined to represent this species in its "pure" form.

Five pairs of accessions were identified with identical DNA profiles, all of which were originally labeled as *M. sieversii*. PI 657760 (DM 34) and PI 657763 (DM 49), both received from a Kyrgyz Republic exploration (Volk et al., 2009a), were identical. Duplicate accession pair PI 650966 and PI 650977 originated from a *M. sieversii* wild-collected seedlot(s) from Site 9 raised at the University of Minnesota and provided back to the NPGS apple collection. PI 614000 and PI 657764, collected from Kazakhstan Sites 9.02 and 9.04, respectively, were both identified as 'Rosmarina Bianca', an old Italian cultivar available from the United Kingdom's National Fruit Collection. The duplicate pair of PI 613953 and PI 613978 was identified as *M. sieversii* with a domestic component. PI 613978 was collected in 1995 from Site 9.05, and a collection note by P. Forsline in 1996 suggested that PI 613953 might be the same tree. The duplicate pair of PI 657072 and PI 657117 was pure *M. sieversii*, although the two accessions were collected from different sites in Kazakhstan (Sites 6 and 12, respectively).

Of the thirty-six *M. sieversii*-labeled accessions classified as "elite" in the NPGS collection, five were determined to be *M. domestica*, 21 *M. sieversii-M. domestica* hybrids, three *M. sieversii* with a *M. domestica* component, and only seven were pure *M. sieversii*. The accession named 'FORM 35' (PI 613967), which was selected in Kazakhstan by Dr. Dzhangaliev and presumed to be *M. sieversii* (GRIN-Global, 2022), was determined to be a *M. sieversii-M. domestica* hybrid, with 'Zigeunerin' as one parent. 'FORM 35' was also determined to be a parent of two other "*M. sieversii*" accessions in the dataset (PI 629319 and PI 629318).

The extent of admixture in sampled populations varied across the original collection sites (Figure 2). All but one accession examined from Site 6, in the Karatau region, were determined to be all pure *M. sieversii*, but only 27% of the accessions of Site 11, also in the Karatau region, were pure *M. sieversii*. A large proportion of the accessions examined from Site 12 (53%) were *M. domestica* cultivars that had been originally considered to be *M. sieversii*. For the site with the largest representation in the dataset, Site 9, located in the Tarbagatai region, 93% of the tested individuals were identified as hybrid or admixed and only 6% were determined to be pure *M. sieversii* (Table 1; Figure 2). The majority (96%) of the 24 sampled accessions originating from outside of Kazakhstan were identified as pure *M. sieversii* (Table 1). Most (75%) of the 76 accessions from Kazakhstan-sourced seedlots that were provided to Dr. James Luby in 2007 that were grown, evaluated, selected among, and then returned to the NPGS were admixed with *M. domestica* (Table S1).

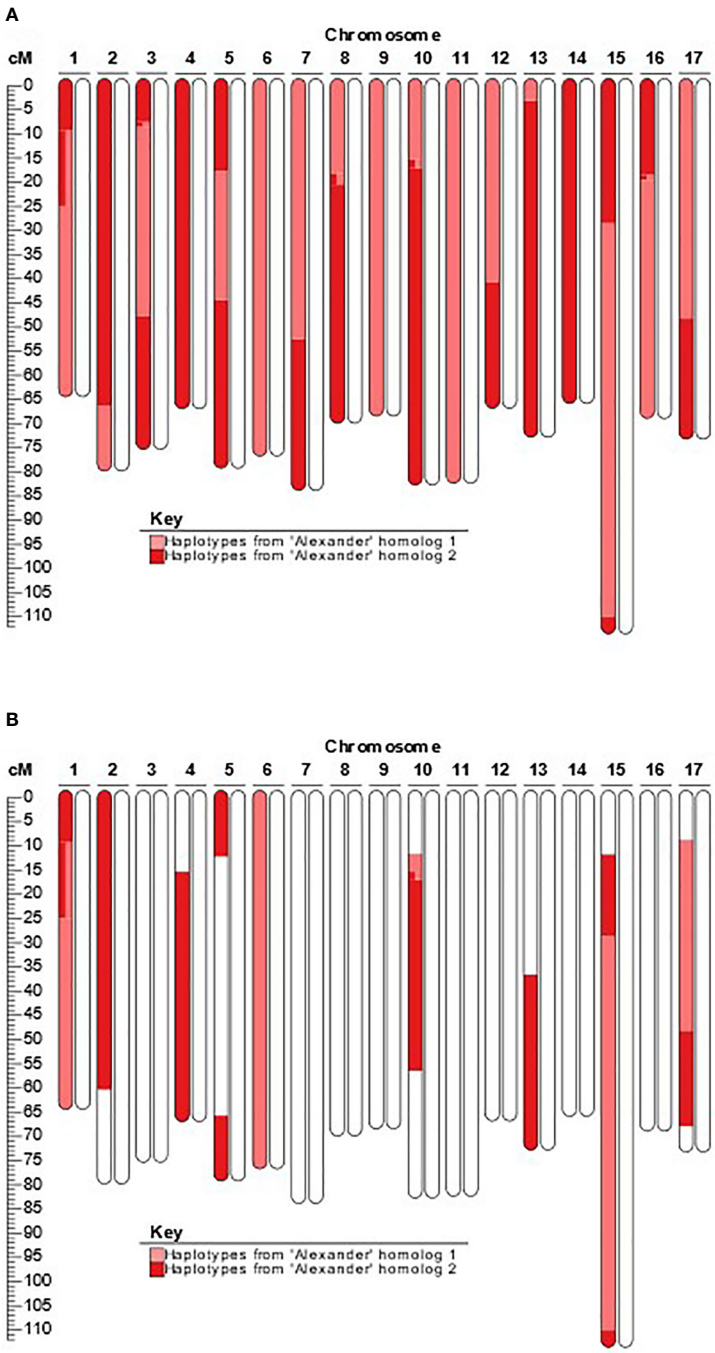


FIGURE 1
Example of newly detected presence of recent *M. domestica* ancestry in accessions previously labeled as pure wild progenitor species. **(A)**. Extended shared haplotypes of 'Alexander' present in the *M. domestica*-*M. sieversii* hybrid accession PI 613979; **(B)** Extended shared haplotypes of 'Alexander' present in the *M. domestica*-*M. sieversii* hybrid accession PI 650959.

Some *M. domestica* cultivars were repeatedly identified as parents and/or grandparents of "*M. sieversii*" accessions (Tables 2, S1, S3). Russian cultivars were identified as the most common source of *M. domestica* contamination in NPGS "*M. sieversii*" accessions collected directly from Tien Shan forests (Sites 3, 4, 5, 9, 11, 12; Figure 2; Table S1) and also as parents and grandparents of "*M. sieversii*" seedlings. 'Alexander', 'Gold Reinette', and 'Charlamoff' were the most prolific grandparents

10.3389/fpls.2022.1015658

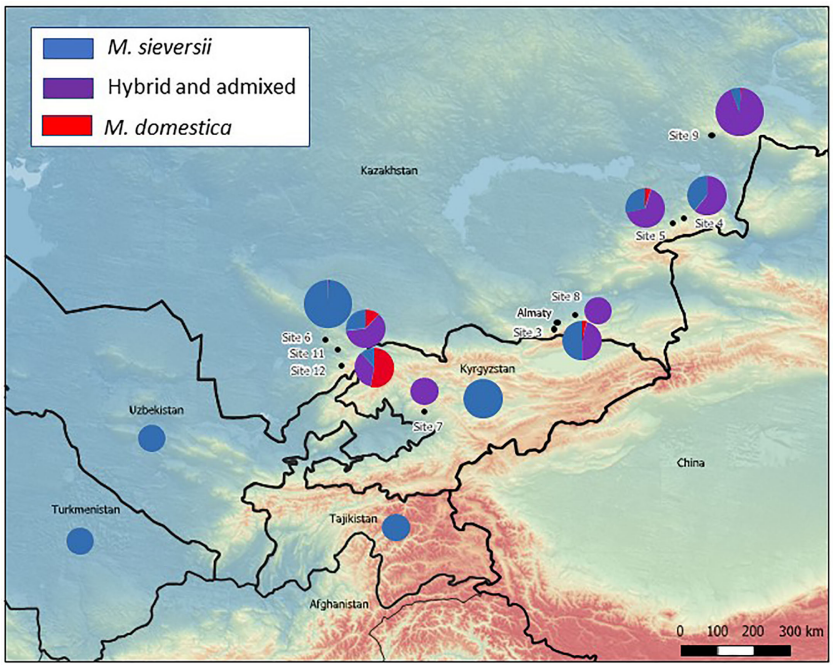


FIGURE 2
Map of *M. sieversii* collection sites for NPGS apple collection genotyped with a 20K SNP array. The subsequently determined proportions of pure *M. sieversii*, pure *M. domestica*, and hybrid/admixed individuals investigated from each site are overlaid as pie charts (small pie chart: 1-10 individuals; medium pie chart: 11-50 individuals; large pie chart: 51-135 individuals).

TABLE 1 Species compositions determined for NPGS *Malus sieversii* collection sites representing various regions in Central Asia.

Site/source (region)	Number of accessions (n)				Proportion of accessions (%)		
	Pure *M. sieversii*	Hybrid/admixed *M. sieversii*	Pure *M. domestica*	Total	Pure *M. sieversii*	Hybrid/admixed *M. sieversii*	Pure *M. domestica*
Site 1 (Tajikistan)	4	0	0	4	100	0	0
Site 2 (Uzbekistan)	6	0	0	6	100	0	0
Site 3 (Zailisky-Almaty, Kazakhstan)	14	13	1	28	50	46	4
Site 4 (Djungarsky-Topelevka, Kazakhstan)	7	11	0	18	39	61	0
Site 5 (Djungarsky-Lepsinsk, Kazakhstan)	12	28	2	42	29	67	5
Site 6 (Karatau, Kazakhstan)	74	1	0	75	99	1	0
Site 7 (Kyrgyzstan)	0	1	0	1	0	100	0
Site 8 (botanic garden, Kazakhstan)	0	1	0	1	0	100	0
Site 9 (Tarbagatai, Kazakhstan)	8	126	1	135	6	93	1
Site 11 (Karatau, Kazakhstan)	7	16	3	26	27	62	12
Site 12 (Talasky, Kazakhstan)	2	6	9	17	12	35	53
Diane Miller donation (Kyrgyzstan)	12	0	0	12	100	0	0
Donation (Turkmenistan)	1	0	0	1	100	0	0

(Table 2). Some non-Russian cultivars were also found directly in the Kazakh landscape or as parents or grandparents of seedlings, such as 'Rosmarina Bianca', an old Italian cultivar (found twice at Site 12 and many times as a recent ancestor at Sites 11 and 12), 'King of the Pippins', originally from England, and cultivars from many other geographical origins.

3.1.2 *Malus orientalis*

Admixture with *M. domestica* was observed in *M. orientalis* NPGS accessions to a lesser extent than in *M. sieversii* accessions. Twenty-six of the 36 accessions labeled as "*M. orientalis*" were pure *M. orientalis*, one (PI 644252) was determined to be fully *M. domestica* (an offspring of 'Golden Delicious' and 'Delicious'), and nine were hybrid/admixed. Of the latter, three were *M. orientalis-M. domestica* hybrids ('Delicious' and 'Eierapfel' were parents), one was a *M. domestica*-exotic hybrid, four were *M. orientalis* with *M. domestica* components, one was a *M. orientalis-M. sieversii* hybrid, One of the *M. domestica-M. orientalis* hybrids, PI 682808.s, was a triploid that shared an allele at every locus with 'Kasseler Renette' (Table S1). In all, 72% of the *M. orientalis* accessions were identified as pure species representatives.

3.1.3 *Malus sylvestris*

The least admixture was detected for *M. sylvestris* NPGS accessions. Of the 44 accessions labeled as *M. sylvestris*, 35 were pure *M. sylvestris*. All *M. sylvestris* accessions from the three half-sib groups had recorded pedigrees that were confirmed with 20K SNP array genotyping and enabled successful whole-genome imputation of the ungenotyped *M. sylvestris* parents 'Oelsen 5' (95.5% of all alleles) and 'Klipphausen' (99.8%). Three accessions labeled as *M. sylvestris* were *M. domestica*, three accessions were *M. sylvestris-M. domestica* hybrids and three were *M. sylvestris* with domestic components (Table S1). In all, 80% of *M. sylvestris* accessions were deemed pure species representatives.

3.2 Phenotypic analysis

Fruit of *M. sieversii-M. domestica* hybrids and *M. orientalis-M. domestica* hybrids were significantly larger than those of their pure wild species counterparts (Table 3). There were too few *M. sylvestris-M. domestica* hybrids to establish that their fruit were significantly larger than those of pure *M. sylvestris* fruit (Table 3). Fruit ground color and proportion of red overcolor were generally similar among counterparts, considering that some fruit might have been sampled and imaged while immature. Pure *M. sieversii* and *M. sieversii-M. domestica* hybrids showed the greatest fruit shape diversity, while fruit of pure *M. orientalis* and *M. sylvestris* accessions were mostly globose and flat-globose (Figures S3–S5; Tables 3, S4). A series of image examples of the relationships between pure *M. sieversii* from Site 12, an *M. sieversii-M. domestica* 'King of the Pippins' hybrid from Site 12, and the cultivar 'King of the Pippins' are shown (Figure 3). In addition, a series of three pure *M. sieversii* accessions, three *M. sieversii-M. domestica* 'Rosmarina Bianca' hybrids, and two Kazakhstan-collected accessions that match genotypes of 'Rosmarina Bianca' are shown (Figure 4).

4 Discussion

It is critical to have access to pure *M. sieversii*, *M. orientalis*, and *M. sylvestris* genebank materials because these species are recognized as valuable to and are being used in costly long-term breeding programs, assessments of domestication, and genetic dissection of traits of interest (Volk et al., 2015a). The extent of hybridization and admixture identified in this study between *M. domestica* and *M. sieversii*, *M. orientalis*, or *M. sylvestris* in the NPGS apple collection is significant. The identification of hybrid/admixed accessions in the NPGS apple collection presented herein may affect findings of research from around the world that used the collection to identify novel alleles and assess genetic relationships.

Pure cultivars and species-*M. domestica* hybrids were identified in NPGS accessions that were labeled as *M. sieversii*, *M. orientalis*, and *M. sylvestris*. In each case, *M. domestica* cultivars in the landscape could have been inadvertently sampled by plant exploration teams. In addition, hybrid trees may have grown from seeds derived from natural pollination occurring between wild species and locally grown *M. domestica* cultivars. Furthermore, collection teams may have sampled fruit from wild trees with seeds resulting from crosses with pollen from cultivars (or hybrids). The following exemplifies how trees in wild populations could have recent cultivar ancestors such as 'Alexander', which was detected as a common recent ancestor of many "*M. sieversii*" accessions in this study. 'Alexander', which was originally from Poland and brought to Kazakhstan in 1865 under the name of 'Aport', was a common cultivar in orchards in Kazakhstan until devastating freezes between 1951 and 1955 that destroyed most orchards. 'Aport' was then replanted within the *M. sieversii* forests (Barbera et al., 2016).

Our results revealed a greater extent of hybridization and admixture in the NPGS apple collection than previously described. Gross et al. (2012) used microsatellite marker genotyping to identify hybrids in the NPGS collection and reported that 10% of the sampled *M. sieversii* and *M. orientalis* and 20% of the sampled *M. sylvestris* was hybrid or admixed, as revealed by STRUCTURE results. Omasheva et al. (2017) reported the genotyping with microsatellite markers of 311 *M. sieversii* trees from 12 wild populations and 16 wild apple clones selected by Dzhangaliev and found the lowest levels of *M. domestica-M. sieversii* admixture in the Kazakh regions of Krutoe truct (89% pure *M. sieversii*) and Tauturgen (92% pure *M. sieversii*). These two sites are located near the present study's

TABLE 2 Named cultivar ancestors of NPGS apple accessions labeled as *Malus sieversii* (excluding genotypic duplicates) identified by pedigree reconstruction using SNP array genotypic information.

Cultivar or accession name	Number of accessions (n)	
	Parent	Grandparent
Anis Aliy	0	1
Charlamoff[1]	4	30
Cheal's Weeping	1	0
Duchess Favorite	0	1
Englische Spitalrenette	0	1
Form 35	3	0
Gold Reinette[1]	4	47
Grågylling	0	1
Alexander[3]	8	51
Kantil Sinap	2	0
King of the Pippins	7	10
Köstlicher	1	6
Kulon Kitaika	2	5
Landsberger Reinette[3]	1	0
Suislepper	5	2
Passe-Pomme Rouge	0	1
Reinette de Hollande[2]	1	7
Reinette Simirenko	2	0
Rosmarina Bianca	7	11
Red Astrachan	1	3
Saint Germain	0	3
Sipolins	0	4
Spasovka Kvasna	1	7
Yellow Bellflower	0	5
Yellow Transparent	0	6
Zigeunerin	2	13

[1]Charlamoff is a parent of Gold Reinette.
[2]Reinette de Hollande is a parent of King of the Pippins.
[3]Alexander is a parent of Landsberger Reinette.

Sites 5 (Djungarsky-Lepsinsk) and 3 (Zailisky-Almaty), respectively, where we detected high levels of hybridization/admixture. The difference could reflect the greater resolution of SNP arrays to detect hybridization/admixture compared to microsatellite marker systems or differences in localized areas of *M. domestica* contamination among sampled locations. The 16 Dzhangaliev clones sampled by Omasheva et al. (2017) were reported to all have some *M. domestica* admixture, which is similar to what we observed with the Dzhangaliev-selected accession 'FORM 35'. Kazakhstan Site 6 had almost no detected admixture and could serve as a source of pure *M. sieversii* for reforestation purposes.

We found that the extent of *M. sieversii* admixture was greater in Kazakhstan than in other sampled Central Asian countries. Ha et al. (2021) also measured admixture of *M. sieversii* and *M. domestica* in Kazakhstan using microsatellite markers. That study sampled 84 *M. sieversii* trees from regions in the east that were near Sites 4 and 5 and towards the west near Site 3 of the current study. All of the populations sampled by Ha et al. (2021) exhibited some hybridization between *M. sieversii* and *M. domestica*, which is consistent with results in the present work for populations sampled from those locations. In contrast, the lower levels of detected admixture in *M. sieversii* sampled from outside of Kazakhstan (Figure 1) may reflect different cultural practices across the landscape resulting in fewer numbers of *M. domestica* trees in or nearby the native stands of *M. sieversii* in other Central Asian countries.

Core sets seek to capture the diversity of populations using a limited number of individuals; the current work has revealed that formerly proposed core sets included admixed individuals. The *M. sieversii* core sets were developed based on microsatellite genotypic data and fruit and disease resistance phenotypic data for individuals from Sites 6 and 9 in Kazakhstan (Volk et al., 2005). Site 6, which was mostly pure *M. sieversii* based on the SNP array analysis, had 31 of 35 core-set individuals that were pure *M. sieversii* and four that were not in the SNP array dataset. In contrast, Site 9, which exhibited high levels of *M. sieversii-M. domestica* admixture in the SNP array analysis, had only four pure *M. sieversii* accessions but 17 *M. sieversii-M. domestica* hybrids, two *M. sieversii* with *M. domestica* components, one *M. domestica*, and no data available for 11 of the 35 core set individuals. A third *M. sieversii* core set was proposed by Richards et al. (2009a) to represent NPGS *M. sieversii* individuals from other collection sites in Kazakhstan. Our results revealed that this set of 35 individuals had 12 pure *M. sieversii*, 19 *M. sieversii-M. domestica* hybrids, three *M. domestica*, and one with no data available. Therefore, the core collections based on Site 9 and the other Kazakhstan sites were identified here as containing a large extent of hybrids and admixture. While the *M. sieversii* core sets chosen by Volk et al. (2005) appear to be have captured the general degree of admixture that is present in the collection sites, the detected contamination with *M. domestica* indicates that these core sets should not be considered as representative of *M. sieversii* diversity there. In any case, each of the three core sets should contain enough representation by *M. sieversii* to enable the discovery of novel alleles.

The results revealed admixture in the individuals represented by the NPGS *M. orientalis* core set individuals proposed in Volk et al. (2009b). Fourteen of the 27 *M. orientalis* core set individuals overlapped with the current SNP dataset. Of those, nine were pure *M. orientalis*, one was a *M. orientalis-M. domestica* hybrid, three were *M. orientalis* with *M. domestica* components, and one was *M. sieversii* with *M. domestica* components. Pure *M. orientalis* was identified in trees that originated from Russia, Turkey, Armenia, and Georgia. Thus, despite contamination with other species, the *M. orientalis* core set appears to be mostly *M. orientalis* and should contain novel alleles. But it should not be considered as representative of *M. orientalis* diversity.

TABLE 3 Some fruit trait observations for pure and hybrid *M. sieversii*, *M. orientalis*, and *M. sylvestris* accessions in the USDA-National Plant Germplasm System apple collection.

Species classification *via* SNP analysis	n	Fruit diameter (cm)	Fruit ground color			Fruit overcolor		Fruit shape							
			green (%)	yellow (%)	red (%)	% Accessions with fruit exhibiting red overcolor	Average % of fruit surface with red overcolor	conical (%)	ellipsoid (%)	ellipsoid-conical (%)	flat (%)	flat-globose (%)	globose (%)	globose-conical (%)	oblong (%)
M. sieversii	87	4.63 ± 0.07 b	86	13	1	66	65	1	1	1	2	25	56	5	8
M. orientalis	18	3.38 ± 0.06 c	50	50		61	65				11	67	22		
M. sylvestris	15	2.94 ± 0.15 c	47	53		27	27					60	40		
M. sieversii × *M. orientalis*	1	4.96 abc	100			100	100						100		
M. sieversii × *M. domestica*	102	5.49 ± 0.11 a	63	36	1	75	75	3			3	28	58	2	6
M. orientalis × *M. domestica*	6	5.45 ± 0.57 ab	83	17		67	57					33	67		
M. sylvestris × *M. domestica*	5	4.62 ± 0.63 abc	60	20	20	80	100					80			20

10.3389/fpls.2022.1015658

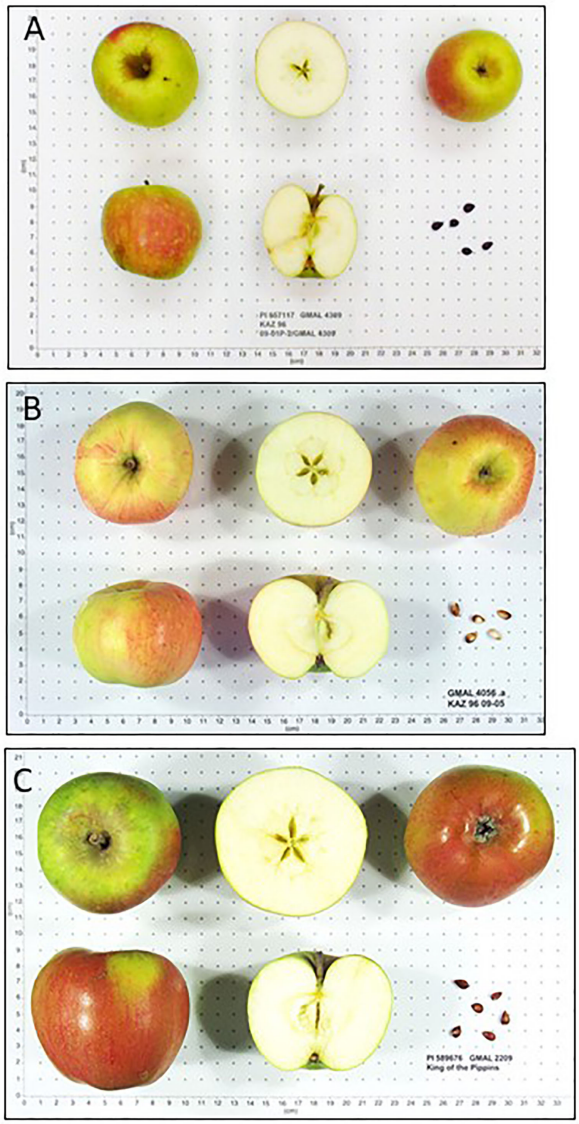


FIGURE 3
Malus fruit images from the USDA-NPGS apple collection **(A)** PI 657117 collected from Kazakhstan Site 12, determined to be pure *M. sieversii*; **(B)** PI 682787 collected from Kazakhstan Site 12, determined to be a hybrid between *M. sieversii* and *M. domestica* 'King of the Pippins'; **(C)** Accession PI 589676, *M. domestica* 'King of the Pippins'.

Previously reported phenotypic assessments performed using NPGS-derived materials are now identified as having unintentionally included hybrid and admixed accessions. Watts et al. (2021) provided phenotypic data for the Apple Biodiversity Collection in Nova Scotia, Canada, derived in part from NPGS accessions. The supplementary data in Watts et al. (2021) has 78 accessions labeled as *M. sieversii*, 55 of which were included in the present study. Of those, two were determined to be *M. domestica*, 21 were *M. sieversii-M. domestica* hybrids, one was *M. domestica* with an exotic component, one was *M. sieversii* with an exotic component, three were *M. sieversii* with *M. domestica* components, one was *M. sieversii* with possible admixture, and only 26 were pure *M. sieversii*. Davies et al. (2022) described the phenotypic divergence between *M. domestica* and *M. sieversii* using the same dataset. While that study revealed significant differences for traits including soluble solids content, bitterness, and firmness during storage between the trees labeled as *M. domestica* and those labeled as *M. sieversii*, those differences could be even more pronounced if only pure species accessions are considered.

The result of extensive *M. domestica* hybridization and admixture in *M. sieversii* accessions might have influenced

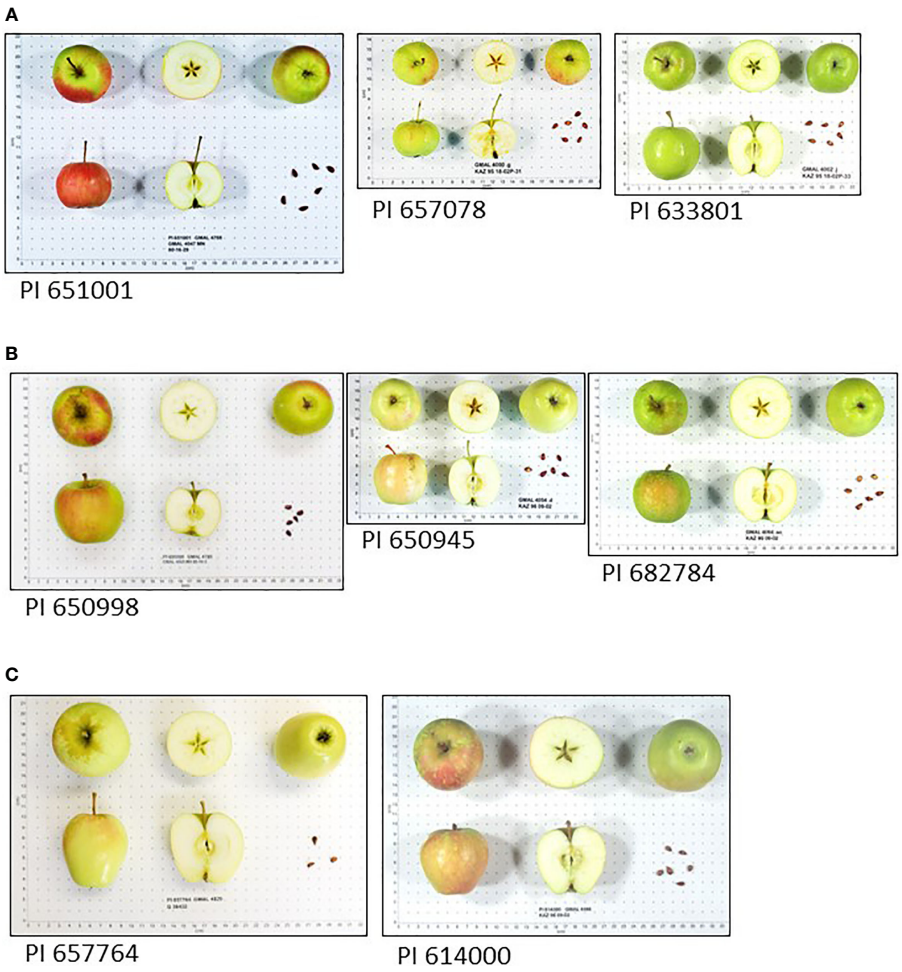


FIGURE 4
Malus fruit images from the USDA-NPGS apple collection **(A)** Three accessions collected in Kazakhstan determined to be pure *M. sieversii*; **(B)** Three accessions collected in Kazakhstan determined to be *M. sieversii-M. domestica* hybrids with 'Rosmarina Bianca' as a parent; **(C)** Two accessions collected as *M. sieversii* in Kazakhstan with genotype matching the United Kingdom National Fruit Collection accession 1951197, 'Rosmarina Bianca'.

previous conclusions of the implied species origins of interesting trait locus alleles. The mapping populations GMAL 4590 and GMAL 4595, used for fine-mapping of the *Ma* locus (Xu et al., 2012), were derived from "*M. sieversii*" parents PI 613971 and PI 613988, respectively, determined here to both be *M. sieversii-M. domestica* hybrids, with the *M. domestica* cultivar Charlamoff being a parent of PI 613971. Four of nine "*M. sieversii*" accessions with PI numbers that were phenotyped for fire blight responses in Washington and West Virginia and often exhibiting resistance (Harshman et al., 2017) were determined here to not be pure *M. sieversii*. Similarly, 21 misidentified and nine pure species PI accessions from the NPGS apple collection were tested for resistance levels to blue mold considering them to all be pure *M. sieversii* (Janisiewicz et al., 2008), and several accessions exhibiting moderate levels of resistance were hybrids and *M. domestica* in our results. Although Wedger et al. (2021) used NPGS apple collection materials that were considered pure according to previous work (Gross et al., 2012), seven of the 15 *M. sieversii* individuals that study sampled from the NPGS were not pure species representatives according to the present analysis. In contrast, mapping population GMAL 4593 was created with parent PI 613981 (Desnoues et al., 2018), a *M. sieversii* pure species representative. Accessions PI 613981 (*M. sieversii*) and PI 633825 (*M. sylvestris*) used by Sun et al. (2020) and Luo et al. (2020) were also pure species representatives. Although *M. sieversii-M. domestica* hybrid and admixed individuals were inadvertently used in previous studies for identifying alleles of interest, discovered novel alleles could still be valuable to breeding and research programs. In some cases, the use of these hybrid and admixed individuals in breeding programs might introduce alleles of interest while incorporating fewer disadvantageous attributes from wild species.

NPGS apple collection materials now determined to be hybrids rather than pure species have been used for whole genome sequence-based projects, which could affect the results and conclusions of those studies. For example, Velasco et al. (2010) used ten *M. sieversii* accessions to determine the relative distinction between *M. domestica* and *M. sieversii*, five of which were determined here to be misidentified (GMAL 4054.a and GMAL 4309.d – *M. domestica*; GMAL 3762.g, GMAL 4304.e, and GMAL 4309.c – *M. sieversii-M. domestica* hybrids). Among the 117 diverse *Malus* accessions used by Duan et al. (2017) to obtain genome sequence data, 10 of the 15 *M. sieversii* accessions from Kazakhstan, the only *M. orientalis*, and three of the seven *M. sylvestris* sampled from the NPGS apple collection were not the expected pure species representatives (while SNP array data are unavailable for a further one *M. sieversii* and three *M. sylvestris* in that study), which helps explain some of the unexpected population structure findings in that report. Migicovsky et al. (2021) also assessed genetic relationships among *M. sieversii*, *M. sylvestris*, and *M. domestica* accessions from the NPGS apple genebank collection, although the specific individuals used were not reported. Inclusion of non-pure accessions of these species would be expected to blur genetic relationships or even exacerbate differences.

Wild species accessions in the NPGS apple collection that were identified as hybrids or admixed here have been included in phylogenetic and wild-cultivar relationships investigations (in addition to the previously mentioned study of Duan et al., 2017). For a *Malus* phylogeny based on chloroplast sequencing, Nikiforova et al. (2013) used several NPGS accessions that were not pure *M. sieversii* as expected, but rather *M. sieversii-M. domestica* hybrids, including GMAL 3610, GMAL 3619, and GMAL 3638, as well as PI 594104, which is actually *M. domestica* with an exotic component. The inclusion of these materials might have affected results, particularly the close relationships detected between *M. sieversii* and *M. domestica*. Similarly, Volk et al. (2015b) inadvertently used some "*M. sieversii*" and "*M. sylvestris*" that were determined here to be pure *M. domestica*, *M. sieversii-M. domestica* hybrids, or *M. sieversii* with *M. domestica* components. Accessions of *M. domestica*, *M. sieversii*, *M. orientalis*, *M. sylvestris*, and other species were classified into three haplotype groups based on chloroplast sequences in which the four species could not be differentiated; however, the clarification of accessions species status here does not remedy the situation. Gharghani et al. (2009) relied on NPGS apple collection materials to determine relationships among old Iranian cultivars and wild *Malus* species. One of the two *M. sylvestris*, ten of the 19 *M. sieversii*, and one of the *M. orientalis* of that study were not pure species representatives according to our results.

The NPGS GRIN-Global database has already been partially updated to reflect revised species statuses as determined herein. This effort will continue and will ideally highlight which accessions labeled as *M. sieversii*, *M. orientalis*, and *M. sylvestris* are pure species representatives. This updated information should ensure that species identities for *M. sieversii*, *M. orientalis*, and *M. sylvestris* are correct in the NPGS apple collection. It is recommended that researchers using accessions of these three species received from the NPGS apple collection update species designations.

Data availability statement

The data presented in the study are deposited in the Genome Database for Rosaceae, accession number tfGDR1063. The data are released and are available here: https://www.rosaceae.org/publication_datasets.

Author contributions

GV, CP, and NH conceived the study. NH analyzed and interpreted data. AH performed laboratory analyses and visualized data. All authors wrote and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This research was partially funded by the USDA National Institute of Food and Agriculture Hatch project 1014919, Crop Improvement and Sustainable Production Systems (WSU reference 00011) and by a 2019 USDA Apple Crop Germplasm Evaluation Grant.

Acknowledgments

SNP data for "Eierapfel" (accession number 100011) was shared by the National Genebank for Plant Genetic Resources for Food and Agriculture (PGREL) in Switzerland. The use of trade, firm, or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the United States Department of Agriculture or the Agricultural Research Service of any product or service to the exclusion of others that may be suitable. USDA is an equal opportunity employer and provider.

Conflict of interest

Author NH is employed by Fresh Forward Breeding and Marketing.

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.2022.1015658/full#supplementary-material

References

Amirchakhmaghi, N., Yousefzadeh, H., Hosseinpour, B., Abdollahi, H., and Larsen, B. (2022). Evaluating responses of Caucasian apple (*Malus orientalis*) from Hyrcanian forests to fire blight (*Erwinia amylovora*) using an *in vitro* assay. *J. Crop Improvement*. 36 (6), 789–800. doi: 10.1080/15427528.2021.2012731

Amirchakhmahgi, N., Yousefzadeh, H., Hosseinpour, B., Espahbodi, K., Aldaghi, M., and Cornille, A. (2018). First insight into genetic diversity and population structure of the Caucasian wild apple (*Malus orientalis* Uglitzk.) in the Hyrcanian forest (Iran) and its resistance to apple scab and powdery mildew. *Genet. Resour. Crop Evol.* 65, 1255–1268. doi: 10.1007/s10722-018-0611-z

Barbera, G., Boschiero, P., Latini, L., and Peix, C. (2016). The wild apple forests of the Tien Shan. International Carlo scarpa prize for gardens in 2016. Fondazione Benetton Studi Ricerche. *Treviso. Italy* p, 52–53.

Bassett, C. L., Glenn, D. M., Forsline, P. L., Wisniewski, M. E., and Farrell, R. E. (2011). Characterizing water use efficiency and water deficit responses in apple (*Malus* × *domestica* Borkh. and *Malus sieversii* Ledeb.) M. Roem. *HortScience* 46, 1079–1084. doi: 10.21273/HORTSCI.46.8.1079

Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., et al. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (*Malus* × *domestica* Borkh). *PLoS One* 9 (10), e110377. doi: 10.1371/journal.pone.0110377

Bramel, P. J., and Volk, G. (2019). *A global strategy for the conservation and use of apple genetic resources* (Germany: Global Crop Diversity Trust. Bonn). doi: 10.13140/RG.2.2.34072.34562

Buiteveld, J., Koehorst-van Putten, H. J. J., Kodde, L., Laros, I., Tumino, G., Howard, N. P., et al. (2021). Advanced genebank management of genetic resources of European wild apple, *Malus sylvestris*, using genome-wide SNP array data. *Tree Genet. Genomes* 17, 32. doi: 10.1007/s11295-021-01513-y

Cornille, A., Giraud, T., Bellard, C., Tellier, A., Le Cam, B., Smulders, M. J. M., et al. (2013). Postglacial recolonization history of the European crabapple (*Malus sylvestris* Mill.), a wild contributor to the domesticated apple. *Mol. Ecol.* 22, 2249–2263. doi: 10.1111/mec.12231

Cornille, A., Giraud, T., Smulders, M. J. M., Roldán-Ruiz, I., and Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends Genet.* 30 (2), 57–65. doi: 10.1016/j.tig.2013.10.002

Cornille, A., Gladieux, P., Smulders, M. J. M., Roldán-Ruiz, I., Laurens, F., Le Cam, B., et al. (2012). New insight into the history of domesticated apple: Secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS Genet.* 9 (5), e1002703. doi: 10.1371/journal.pgen.1002703

Davies, T., Watts, S., McClure, K., Migicovsky, Z., and Myles, S. (2022). Phenotypic divergence between the cultivated apple (*Malus domestica*) and its primary wild progenitor (*Malus sieversii*). *PLoS One* 17 (3), e0250751. doi: 10.1371/journal.pone.0250751

Denancé, C., Muranty, H., and Durel, C.-E. (2020). MUNQ-*Malus* UniQue genotype code for grouping apple accessions corresponding to a unique genotypic profile, VI edn. *Portail Data INRAE*. doi: 10.15454/HKGMAS

Desnoues, E., Norelli, J. L., Aldwinckle, H. S., Wisniewski, M. E., Evans, K. M., Malnoy, M., et al. (2018). Identification of novel strain-specific and environment-dependent minor QTLs linked to fire blight resistance in apples. *Plant Mol. Biol. Rep.* 36, 247–256. doi: 10.1007/s11105-018-1076-0

Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., et al. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nat. Commun.* 8, 249. doi: 10.1038/s41467-017-00336-7

Fazio, G., Aldwinckle, H. S., Volk, G. M., Richards, C. M., Janisiewicz, W. J., and Forsline, P. L. (2009). Progress in evaluating *Malus sieversii* for disease resistance and horticultural traits. *Acta Hortic.* 814, 59–66. doi: 10.17660/ActaHortic.2009.814.2

Fazio, G., Chao, C. T., Forsline, P. L., Richards, C., and Volk, G. (2014). Tree and root architecture of *Malus sieversii* seedlings for rootstock breeding. *Acta Hortic.* 1058, 585–594. doi: 10.17660/ActaHortic.2014.1058.75

Forsline, P. L., Aldwinckle, H. S., Dickson, E. E., Luby, J. J., and Hokanson, S. C. (2002). Collection, maintenance, characterization, and utilization of wild apples of Central Asia. *Hortic. Rev.* 29, 1–62. doi: 10.1002/9780470650868.ch1

Gharghani, A., Zamani, Z., Talaie, A., Oraguzie, N. C., Fatahi, R., Hajnajari, H., et al. (2009). Genetic identity and relationships of Iranian apple (*Malus* × *domestica* Borkh.) cultivars and landraces, wild *Malus* species and representative old apple cultivars based on simple sequence repeat (SSR) marker analysis. *Genet. Resour. Crop Evol.* 56, 829–842. doi: 10.1007/s10722-008-9404-0

Gross, B. L., Henk, A. D., Forsline, P. L., Richard, C. M., and Volk, G. M. (2012). Identification of interspecific hybrids among domesticated apple and its wild relatives. *Tree Genet. Genomes* 8, 1223–1235. doi: 10.1007/s11295-012-0509-4

Gutierrez, B., Battaglia, K., and Zhong, G.-Y. (2020). Preserving the future with the USDA plant genetic resources unit tart cherry, grape, and apple germplasm collections. *J. Amer. Pomological Soc.* 74, 97–103.

Ha, Y.-H., Oh, S.-H., and Lee, S.-R. (2021). Genetic admixture in the population of wild apple (*Malus sieversii*) from the Tien Shan mountains, Kazakhstan. *Genes* 12, 104. doi: 10.3390/genes12010104

Harshman, J. M., Evans, K. M., Allen, H., Potts, R., Flamenco, J., Aldwinckle, H. S., et al. (2017). Fire blight resistance in wild accessions of *Malus sieversii*. *Plant Dis.* 101, 1738–1745. doi: 10.1094/PDIS-01-17-0077-RE

Howard, N. P., Peace, C., Silverstein, K. A. T., Poets, A., Luby, J. J. , Vanderzande, S., et al. (2021a). The use of shared haplotype length information for pedigree reconstruction in asexually propagated outbreeding crops, demonstrated for apple and sweet cherry. *Hortic. Res.* 8, 202. doi: 10.1038/s41438-021-00637-5

Howard, N. P., Troggio, M., Durel, C.-E., Muranty, H., Denancé, H. , Bianco, L., et al. (2021b). Integration of infinium and axiom SNP array data in the outcrossing species *Malus* × *domestica* and causes for seemingly incompatible calls. *BMC Genomics* 22, 246. doi: 10.1186/s12864-021-07565-7

Howard, N. P., van de Weg, E., Tillman, J., Tong, C. B. S., Silverstein, K. A. T., and Luby, J. J. (2018). Two QTL characterized for soft scald and soggy breakdown in apple (*Malus* × *domestica*) through pedigree-based analysis of a large population of interconnected families. *Tree Genet. Genomes* 14, 2. doi: 10.1007/s11295-017-1216-y

Janisiewicz, W. J., Saftner, R. A., Conway, W. S., and Forsline, P. L. (2008). Preliminary evaluation of apple germplasm from Kazakhstan for resistance to postharvest blue mold in fruit caused by *Penicillium expansum*. *HortScience* 43, 420–426. doi: 10.21273/HORTSCI.43.2.420

Jurick, W. M., Janisiewicz, W. J., Saftner, R. A., Vico, I., Gaskins, V. L., Park, E., et al. (2011). Identification of wild apple germplasm (*Malus* spp.) accessions with resistance to the postharvest decay pathogens *Penicillium expansum* and *Colletotrichum acutatum*. *Plant Breed.* 130, 481–486. doi: 10.1111/j.1439-0523.2011.01849.x

Khadivi, A., Mirheidari, F., Moradi, Y., and Paryan, S. (2020). *Malus orientalis* Uglitzk., an important genetic resource to improve domestic apples: characterization and selection of the promising accessions. *Euphytica* 216, 189. doi: 10.1007/s10681-020-02720-9

Kostick, S. A., and Luby, J. J. (2022). Apple fruit size QTLs on chromosomes 8 and 16 characterized in 'Honeycrisp'-derived germplasm. *Agronomy* 12 (6), 1279. doi: 10.3390/agronomy12061279

Liu, Z., Li, X., Wen, X., Zhang, Y., Ding, Y., Zhang, Y., et al. (2021). PacBio full-length transcriptome of wild apple (*Malus sieversii*) provides insights into canker disease dynamic response. *BMC Genomics* 22, 52. doi: 10.1186/s12864-021-07366-y

Luby, J., Forsline, P., Aldwinckle, H., Bus, V., and Geibel, M. (2001). Silk road apples–collection, evaluation, and utilization of *Malus sieversii* from Central Asia. *HortScience* 36, 225–231. doi: 10.21273/HORTSCI.36.2.225

Luo, F., Norelli, J. J., Howard, N. P., Wisniewski, M., Flachowsky, H., Hanke, M.-V., et al. (2020). Introgressing blue mold resistance into elite apple germplasm by rapid cycling breeding and foreground and background DNA-informed selection. *Tree Genet. Genomes* 15, 28. doi: 10.1007/s11295-020-1419-5

Maguylo, K., and Bassett, C. (2014). Phenotyping *M. sieversii* collections from Kazakhstan for leaf traits and tree architecture. *Acta Hortic.* 1058, 335–341. doi: 10.17660/ActaHortic.2014.1058.40

Migicovsky, Z., Gardner, K. M., Richards, C., Chao, C. T., Schwaninger, H. R., Fazio, G., et al. (2021). Genomic consequences of apple improvement. *Hortic. Res.* 8, 9. doi: 10.1038/s41438-020-00441-7

Migicovsky, Z., and Myles, S. (2017). Exploiting wild relatives for genomics-assisted breeding of perennial crops. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00460

Moradi, Y., Khadivi, A., and Mirheidari, F. (2022). Multivariate analysis of oriental apple (*Malus orientalis* Uglitzk). based on phenotypic and pomological characterizations. *Food Sci. Nutr*. 10, 2532–2541. doi: 10.1002/fsn3.2858

Nikiforova, S. V., Cavalieri, D., Velasco, R., and Goremykin, V. (2013). Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Mol. Biol. Evol.* 30 (8), 1751–1760. doi: 10.1093/molbev/mst092

Norelli, J. L., Wisniewski, M., and Droby, S. (2014). identification of a QTL for postharvest disease resistance to *Penicillium expansum* in *Malus sieversii*. *Acta Hortic.* 1053, 199–203. doi: 10.17660/ActaHortic.2014.1053.21

Norelli, J. J., Wisniewski, M., Fazio, G., Burchard, E., Gutierrez, B., Levin, E., et al. (2017). Genotyping-by-sequencing markers facilitate the identification of quantitative trait loci controlling resistance to *Penicillium expansum* in *Malus Sieversii*. *PLoS One* 12 (3), e0172949. doi: 10.1371/journal.pone.0172949

Nussenov, Z. (2018). The eleventh letter (of Johann Sievers) from the top of Tarbagatai on June 30 [1793]. *Lett. Siberia Almaty* 90, 52–60.

Omasheva, M. Y., Flachowsky, H., Ryabushkina, N. A., Pozharskiy, A. S., Galiakparov, N. N., and Hanke, M.-V. (2017). To what extent do wild apples in Kazakhstan retain their genetic integrity? *Tree Genet. Genomes* 13, 52. doi: 10.1007/s11295-017-1134-z

Peace, C. P., Bianco, L., Troggio, M., van de Weg, E., Howard, N. P., Cornille, A., et al. (2019). Apple whole genome sequences: recent advances and new prospects. *Hortic. Res.* 6, 59. doi: 10.1038/s41438-019-0141-7

R Core Team (2022) *The r project for statistical computing*. Available at: https://www.r-project.org/.

Reim, S., Hölтken, A., and Höfer, M. (2013). Diversity of the European indigenous wild apple (*Malus sylvestris* (L.) Mill.) in the East Ore mountains (Osterzgebirge), Germany: II. *Genetic-characterization Genet. Resour. Crop Evol.* 60, 879–892. doi: 10.1007/s10722-012-9885-8

Reim, S., Lochschmidt, F., Proft, A., and Höfer, M. (2020). Genetic integrity is still maintained in natural populations of the indigenous wild apple species *Malus sylvestris* (Mill.) in Saxony as demonstrated with nuclear SSR and chloroplast DNA markers. *Ecol. Evol.* 10, 11798–11809. doi: 10.1002/ece3.6818

Richards, C. M., Volk, G. M., Reeves, P. A., Reilley, A. A., Henk, A. D. , Forsline, P. L., et al. (2009a). Selection of stratified core sets representing wild apple (*Malus sieversii*). *HortScience* 134, 228–235. doi: 10.21273/JASHS.134.2.228

Richards, C. M., Volk, G. M., Reilley, A. A., Henk, A. D., Lockwood, D. R. , Reeves, P. A., et al. (2009b). Genetic diversity and population structure in *Malus sieversii*, a wild progenitor species of domesticated apple. *Tree Genet. Genomes* 5, 339–347. doi: 10.1007/s11295-008-0190-9

Robinson, J. P., Harris, S. A., and Juniper, B. E. (2001). Taxonomy of the genus *Malus* mill. (Rosaceae) with emphasis on the cultivated apple, *Malus domestica* borkh. *Plant Syst. Evol.* 226, 35–58. doi: 10.1007/s006060170072

Ruhsam, M., Jessop, W., Cornille, A., Renny, J., and Worrell, R. (2019). Crop-to-wild introgression in the European wild apple *Malus sylvestris* in Northern Britain. *Forestry* 92, 85–96. doi: 10.1093/forestry/cpy033

Schnitzler, A., Arnold, C., Cornille, A., Bachmann, O., and Schnitzler, C. (2014). Wild European apple (*Malus sylvestris* (L.) Mill.) population dynamics: Insight from genetics and ecology in the Rhine valley. priorities for a future conservation programme. *Plos One* 9 (5), e96596. doi: 10.1371/journal.pone.0096596

Singh, J., Sun, M., Cannon, S. B., Wu, J., and Khan, A. (2021). An accumulation of genetic variation and selection across the disease-related genes during apple domestication. *Tree Genet. Genomes* 17, 29. doi: 10.1007/s11295-021-01510-1

Sun, X., Jiao, C., H.Chao, C. T., Ma, Y., Duan, N., Khan, A., et al. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat. Genet.* 52, 1423–1432. doi: 10.1038/s41588-020-00723-9

USDA (2022) *GRIN-global*. Available at: https://npgsweb.ars-grin.gov/gringlobal/search (Accessed July 17, 2022).

Vanderzande, S., Howard, N. P., Cai, L., Da Silva Linge, C., Antanaviciute, L., Bink, M. C. A. M., et al. (2019). High-quality, genome-wide SNP genotypic data for pedigreed germplasm of the diploid outbreeding species apple, peach, and sweet cherry through a common workflow. *PLoS One* 14 (6), e0210928. doi: 10.1371/journal.pone.0210928

Van Nocker, S., Berry, G., Najdowski, J., Michelutti, R., Luffman, M., Forsline, P., et al. (2012). Genetic diversity of red-fleshed apples (*Malus*). *Euphytica* 185, 281–293. doi: 10.1007/s10681-011-0579-7

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus* × *domestica* Borkh.). *Nat. Genet.* 42 (10), 833–841. doi: 10.1038/ng.654

Volk, G. M., Chao, C. T., Norelli, J., Brown, S. K., Fazio, G. , Peace, C., et al. (2015a). The vulnerability of US apple (*Malus*) genetic resources. *Genet. Resour. Crop Evol.* 62, 765–794. doi: 10.1007/s10722-014-0194-2

Volk, G. M., Henk, A. D., Baldo, A., Fazio, G., Chao, C. T. , and Richards, C. M. (2015b). Chloroplast heterogeneity and historical admixture within the genus *Malus*. Am. *J. Bot.* 102, 1198–1208. doi: 10.3732/ajb.1500095

Volk, G. M., Richards, C. M., Henk, A. D., Reilley, A., Miller, D. D. , and Forsline, P. L. (2009a). Novel diversity identified in a wild apple population from the Kyrgyz Republic. *HortScience* 44, 516–518. doi: 10.21273/HORTSCI.44.2.516

Volk, G. M., Richards, C. M., Henk, A. D., Reilley, A. A., Reeves, P. A., Forsline, P. L., et al. (2009b). Capturing the diversity of wild *Malus orientalis* from Georgia, Armenia, Russia, and Turkey. *J. Amer. Soc. Hortic. Sci.* 134, 453–459. doi: 10.21273/JASHS.134.4.453

Volk, G. M., Richards, C. M., Reilley, A. A., Henk, A. D., Forsline, P. L., and Aldwinckle, H. S. (2005). Ex situ conservation of vegetatively propagated species: Development of a seed-based core collection for *Malus sieversii*. J. *Amer. Soc Hortic. Sci.* 130, 203–210. doi: 10.21273/JASHS.130.2.203

Wang, A., Aldwinckle, H., Forsline, P., Main, D., Fazio, G., Brown, S., et al. (2012). EST contig-based SSR linkage maps for *Malus* × *domestica* cv royal gala and an apple scab resistant accession of m. sieversii, the progenitor species of domestic apple. *Mol. Breed.* 29, 379–397. doi: 10.1007/s11032-011-9554-1

Watkins, R., and Smith, R. A. (1997). *Apple descriptors* (Rome: International Board for Plant Genetic Resources Secretariat). Available at: https://www.bioversityinternational.org/e-library/publications/detail/apple-descriptors/.

Watts, S., Migicovsky, Z., McClure, K. A., Yu, C. H. J., Amyotte, B., and Baker, T. (2021). Quantifying apple diversity: A phenomic characterization of Canada's apple biodiversity collection. *Plants People Planet* 3, 747–760. doi: 10.1002/ppp3.10211

Wedger, M. J., Schumann, A. C., and Gross, B. L. (2021). Candidate genes and signatures of directional selection on fruit quality traits during apple domestication. *Am. J. Bot.* 108, 616–627. doi: 10.1002/ajb2.1636

Wisniewski, M., Artlip, T., Liu, J., Ma, J., Burchard, E., Norelli, J., et al. (2020). Fox hunting in wild apples: Searching for novel genes in *Malus sieversii*. *Int. J. Mol. Sci.* 21, 9516. doi: 10.3390/ijms21249516

Xu, K., Wang, A., and Brown, S. (2012). Genetic characterization of the *Ma* locus with pH and titratable acidity in apple. *Mol. Breed.* 30, 899–912. doi: 10.1007/s11032-011-9674-7

Zhang, C., Chen, X., He, T., Liu, X., Feng, T., and Yuan, Z. (2007). Genetic structure of *Malus sieversii* population from xianjiang, China, revealed by SSR markers. *J. Genet. Genomics* 34 (10), 947–955. doi: 10.1016/S1673-8527(07)60106-4

frontiers | Frontiers in Plant Science

TYPE Review
PUBLISHED 17 November 2022
DOI 10.3389/fpls.2022.1008904

OPEN ACCESS

EDITED BY
Andrés J. Cortés,
Colombian Corporation for Agricultural Research (AGROSAVIA), Colombia

REVIEWED BY
Paul Gepts,
University of California, Davis, United States
Thomas M. Davis,
University of New Hampshire, United States
Robert Philipp Wagensommer,
Free University of Bozen-Bolzano, Italy

*CORRESPONDENCE
Jacqueline Batley
Jacqueline.batley@uwa.edu.au

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

RECEIVED 01 August 2022
ACCEPTED 25 October 2022
PUBLISHED 17 November 2022

CITATION
Tirnaz S, Zandberg J, Thomas WJW, Marsh J, Edwards D and Batley J (2022) Application of crop wild relatives in modern breeding: An overview of resources, experimental and computational methodologies. *Front. Plant Sci.* 13:1008904. doi: 10.3389/fpls.2022.1008904

Application of crop wild relatives in modern breeding: An overview of resources, experimental and computational methodologies

Soodeh Tirnaz, Jaco Zandberg, William J. W. Thomas, Jacob Marsh, David Edwards and Jacqueline Batley*

School of Biological Sciences, University of Western Australia, Perth, WA, Australia

Global agricultural industries are under pressure to meet the future food demand; however, the existing crop genetic diversity might not be sufficient to meet this expectation. Advances in genome sequencing technologies and availability of reference genomes for over 300 plant species reveals the hidden genetic diversity in crop wild relatives (CWRs), which could have significant impacts in crop improvement. There are many *ex-situ* and *in-situ* resources around the world holding rare and valuable wild species, of which many carry agronomically important traits and it is crucial for users to be aware of their availability. Here we aim to explore the available *ex-/in- situ* resources such as genebanks, botanical gardens, national parks, conservation hotspots and inventories holding CWR accessions. In addition we highlight the advances in availability and use of CWR genomic resources, such as their contribution in pangenome construction and introducing novel genes into crops. We also discuss the potential and challenges of modern breeding experimental approaches (e.g. *de novo* domestication, genome editing and speed breeding) used in CWRs and the use of computational (e.g. machine learning) approaches that could speed up utilization of CWR species in breeding programs towards crop adaptability and yield improvement.

KEYWORDS

pangenome, wild species, modern breeding, ex situ resources, *in situ* resources

What can CWRs offer?

The world population is estimated to come close to 10 billion by 2050, while a food gap of more 50% is expected between 2006 and 2050 (Ranganathan et al., 2016). In addition, the growing consequences of climate change, such as increasing weed prevalence and the occurrence of severe disease epidemics and drought

stresses (Raza et al., 2019) will lead towards billions of dollars of crop yield losses worldwide (Gregory et al., 2009; Mittler and Blumwald, 2010). The IPCC (2014) has projected yield losses of up to 25% due to climate change if crop adaptation and improvement are not implemented (IPCC, 2014). At the same time, diets are changing, with shifting nutritional demands toward gluten free, plant-based protein and low GI (glycaemic index) products (Gaikwad et al., 2020). As a result, there is an urgent need for plant breeders to develop new traits in addition to agronomically important traits such as disease resistance, drought tolerance, and yield improvements. On top of these challenges, the effect of the recent COVID-19 pandemic on future agricultural industries has likely added financial strain to both production and distribution chains due to restricted food trade policies and closure of food production facilities (Aday and Aday, 2020). These factors put farmers in a precarious position, with growing pressure to increase production, while they are placed in an increasingly vulnerable position to crop failure and infrastructure setbacks.

Providing breeders access to diverse genetic resources is essential to facilitate, accelerate and optimise crop improvement approaches while domestication bottlenecks have also restricted modern breeding populations (Allaby et al., 2019). The reduction in genetic diversity induced by domestication bottleneck is well documented among many crops such as common bean (Gepts et al., 1986; Papa and Gepts, 2003). Compared to the domesticated population, there are tremendous genetic diversity persists among crop wild relatives (CWRs). The structure of genetic diversity among wild populations appears to be stronger than domesticated; for example in common bean, the diversity of domesticated beans showed limited geographical structure and much less differentiation among populations and regions while in wild bean population even geographically-short-distanced populations carry significant genetic diversity (Papa and Gepts, 2003). As a result, the addition of CWRs to the current breeding programs can significantly widen the source of genetic variation and selection towards yield, resistance and nutritional quality improvement in crops. CWRs can be defined as any taxon belonging to the same genus as a crop; however this definition will include species that are both closely or remotely related to crops (Maxted et al., 2006). In a narrower definition CWRs belong to the same genus of the crop and are closely related to the crops (i.e they are ranked as same the species or same subgenus) (Maxted et al., 2006; Perrino and Perrino, 2020). Advances in breeding techniques, such as genome sequencing, pangenome construction and *de novo* domestication, have been facilitating traits/gene selection from both closely and remotely, related species where fertility and compatibility will be a barrier in traditional breeding approaches, related CWRs to crops. There are a number of successful examples of CWRs application in breeding, such as disease and pest resistance improvement in wheat, rice, potato, tomato, cassava, sunflower, banana and lettuce; yield improvement in wheat and rice; and improving tolerance to abiotic stress in rice, tomato, barley and chickpea (Hajjar and Hodgkin, 2007). CWRs have also contributed beneficial traits related to ideal plant architecture and weed suppression in rice (Inagaki et al., 2021).

The diversity among CWRs could also be used to decrease the rate of gene/genetic erosion, which has been happening over decades of crop domestication and intense breeding (Schouten et al., 2019). The FAO estimates that ~75% of the genetic diversity in crop varieties has been lost over the past century (FAO, 1999; Khoury et al., 2022). Genetic erosion restricts breeders by limiting sources of selection for identifying desirable agronomic traits. For instance, 96% of peas grown in the US originated from only 9 varieties (Esquinas-Alcázar, 2005). This limited genetic pool will significantly decrease diversity for natural and artificial selection, and intensify the vulnerability of modified varieties to rapid climate changes and new environmental stresses (Esquinas-Alcázar, 2005). Pangenomic analyses in soybean also revealed a reduction in mean gene count per individual due to domestication (Bayer et al., 2022), with disproportionately high levels of biotic and abiotic stress genes lost in modern breeding populations compared to CWRs (Liu et al., 2020). Fortunately, the application of wild species in breeding programs can be used to recover lost diversity caused by erosion, and boost diversity among the crops. SNP array analysis showed that genetic diversity among commercial tomato varieties (from NW Europe) increased by a factor of eight over 7 decades (starting from the 1950s) as a result of the introgression of many disease resistances genes from wild relatives (Schouten et al., 2019).

The application of CWRs in breeding has been also shown to deliver huge economic returns in agricultural industries worldwide, with their annual contribution to the world economy estimated at around US $186.3 billion in 2020 (Tyack et al., 2020; Bohra et al., 2022). It has been estimated that around 30% of crop yield improvement since 1945, valued worldwide at around US $100 billion, is a result of CWR use in crop breeding (Pimentel et al., 1997; Brozynska et al., 2016). In tomato, one wild variety provided genes increasing solids content by 2.4% which was worth US$250 million a year to the global tomato industry; and genes from three wild peanut varieties increased resistance to the root knot nematode, for potential savings of around US $100 million each year worldwide (Maxted, 2008).

Despite all the potential that CWRs can offer to improve breeding programs, their *in-situ* (in their natural habitats) and *ex-situ* (outside their natural habitats) conservation has been neglected over many years, leading to their potential extinction. Global and local studies have been conducted to guide CWR conservation strategies and estimate the potential loss of diversity of CWRs if the required actions have not been taken. In the US, conservation assessments for 600 CWRs show 42 taxa

(7%) are critically endangered in their natural habitats, 297 (50%) are endangered, 166 (28%) are vulnerable, 66 (11%) are near threatened, and only 23 (3%) are of least concern (Khoury Colin et al., 2020). Another CWR conservation study revealed that the diversity of CWRs is poorly represented in genebanks while out of 1,076 taxa related to 81 crops, for 313 (29%) taxa no germplasm accessions exist, and for 257 (23%) taxa fewer than ten accessions exist (Castañeda-Álvarez et al., 2016). A conservation study on 29 threatened CWRs in Italy, also indicates 23 out of 29 species, have no gene pool at all. In addition, there is not enough data of their *ex-situ* and *in-situ* conservation while 16 and 22 species were identified as high priority for *ex-situ* and *in-situ* conservation respectively (Perrino and Wagensommer, 2022).

Rapid advancements in sequencing technology and computational approaches offer excellent opportunities to fully harness CWR diversity for crop improvement. However, the availability and accessibility of the existing CWR genebank and germplasm resources, capability of modern breeding methodologies and techniques in use of CWRs conservation strategies are currently not well developed to support their full potential and contribution in the current breeding programs. In this regard, here we discuss available in-/*ex-situ* resources for the preservation of CWR variation and the advances in the modern experimental methodologies and computational tools to facilitate capturing the genetic diversity among CWR and their utilization in breeding.

Ex-situ resources

Ex-situ resources, e.g. genebanks and botanical gardens, facilitate user access to plant samples without the need for collecting samples directly from their natural habitat, which can be laborious and complicated when species only exist in remote locations and in most cases need collecting permit (PolicyReport, 2016) and in many cases may not accessible because of political or socio-economic unrest. The number of accessions held worldwide in genebanks estimated at ~7.4 million accessions in 2009, which increased more than 1.4 million from 1996, ~30% of this increase associated with CWR (van Bemmelen van der Plaat et al., 2021). There are now more than 1750 genebanks worldwide, with 130 of them holding more than 10,000 accessions each (Bohra et al., 2021). Wheat (856,168 accessions), rice (773,948 accessions), barley (466,531 accessions), maize (327,932 accessions) and bean (261,963 accessions) are the most represented crops across the world's genebanks (Wambugu et al., 2018).

To facilitate global access and the conservation of genetic diversity of cultivated and CWR species, genebanks work collaboratively; for instance, Genesys is a database (platform) that contains information of around 4 million accessions across 450 institutes and allows researchers, breeders and policymakers to browse across all genebanks (https://www.genesys-pgr.org/content/about/about) (Table 1). The Genesys database also includes accession information of three of the world's largest genebank databases; the Consultative Group on International Agricultural Research (CGIAR), European Search Catalogue for Plant Genetic Resources (EURISCO), and the U.S. National Plant Germplasm System (NPGS). In contrast to CGIAR and EURISCO that hold both crops and CWRs accessions, the NPGS collection mainly focuses on crop germplasm (https://www.ars-grin.gov/Pages/Collections#bkmk-1). The EURISCO database contains over 2 million accessions of crop plants and their wild relatives preserved ex situ by about 400 institutes (https://eurisco.ipk-gatersleben.de/apex/eurisco_ws/r/eurisco/home). CGIAR is a partnership of 11 genebanks conserving over 700,000 accessions of cereals, grain legumes, forages, tree species, root and tuber crops and banana and their wild relatives (Table 1). For instance, one of the CGIAR genebank partners is the International Institute of Tropical Agriculture (IITA) which holds over 28,000 accessions of plant material or germplasm of major African crops, including cassava, plantain and banana, yam, soybean, bambara ground-nut and maize. IITA holds the world's largest collection of cowpeas, with 15,1222 samples from 88 countries, representing almost half of the global diversity (https://www.iita.org/research/genetic-resources/). There are also several genebanks that hold local genetic diversity of crop wild relatives, for example, the Karlsruher Institute of Technology (KIT) collected around 250 species of CWRs with 4500 accessions from all over Germany (https://www.botanik.kit.edu/garten/english/1056.php) (Table 1).

Recourses available in genebanks have been used in a number of studies, for example Abdallah et al. (2020) obtained 285 accessions, representing 13 *Lathyrus* (grass pea) species, from The International Center for Agricultural Research in the Dry Areas (ICARDA) and showed that wild *Lathyrus* species have higher resistance to broomrape weeds (*Orobanche* spp.), a root holoparasitic plant that causes significant damage to legume crops (Abdallah et al., 2021). Dida et al. (2021) obtained 52 finger millet accessions, including landraces, wild lines and hybrids between wild and cultivated genotypes, from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) and Genetic Resources Research Institute (GeRRI) genebanks and found that wild accessions were more resistant to blast disease, caused by the *Magnaporthe grisea* fungus, in comparison to the cultivated accessions (Dida et al., 2021).

In addition to the germplasm conservation, there are also genebanks that provide seed kits to smallholder farmers to improve local access to the crop diversity towards better nutrition and supporting climate-resilient agriculture these also assist with the improvement of local genetic diversity among crops. For example, the World Vegetable Center (WorldVeg) genebank distributed over 42,000 seed kits, containing over 183,000 vegetable seeds, to smallholder farmers in Tanzania, Kenya and Uganda, between 2013 and 2017. The kits contained seed of promising accessions and open-

TABLE 1 A list of well-known resources and platforms that include CWR species.

Type	Platform/ resource	Details	Reference/Access
Ex situ: Genebanks	KIT	4500 accessions of CWR across Germany	https://www.botanik.kit.edu/garten/english/1056.ph
	APG	70,000 accessions of pasture and forage	https://pir.sa.gov.au/research/australian_pastures_genebank
	World Vegetable Center	Holding 64,948 accessions which represent 330 species from 155 countries. Species are globally important vegetables such as tomato, onion, peppers, and cabbage as well as more than 10,000 accessions of traditional vegetables	https://avrdc.org/
	Genesys	Partnership with the 3 following genebanks	https://www.genesys-pgr.org/
	&Eurisco	Includes hundreds of research centers, genebanks and institutions across Europe	https://eurisco.ipk-gatersleben.de/apex/f?p=103:1
	&NPGS	Includes information of genebanks across U.S., managed by United States Department of Agriculture (USDA). NPGS comprises 20 separate institutions involved in plant germplasm collection, preservation, and distribution.	https://www.ars-grin.gov/Pages/Collections
	&CGIAR	Parentship with the following 11 genebanks	
	&African Rice Center	Holding almost 22,000 rice accessions, 85% of which originated in Africa	https://www.genebanks.org/genebanks/africarice/
	&Bioversity International	Holding the world's largest collection of banana diversity, including more than 1,500 accessions of cultivated and wild species	https://www.genebanks.org/genebanks/biodiversity-international/
	&CIAT	Holding diverse collections of beans and tropical forages as seed and whole plants, and cassava *in vitro* and as small plants	https://www.genebanks.org/genebanks/ciat/
	&CIMMYT	Holding large collections of maize (28,000 accessions), including wild teosinte and Tripsacum wild relatives of maize and large collection of wheat (150,000 accessions) including landraces and wild relatives.	https://www.genebanks.org/genebanks/cimmyt/
	&CIP	The world's largest collection of Potato and sweet potato and contains nearly all of the potato wild relatives	https://www.genebanks.org/genebanks/international-potato-centre/
	&ICARDA	Holding diverse collections of barley and wheat, grain legumes and forages, mostly traditional landraces and wild species from the Fertile Crescent	https://www.genebanks.org/genebanks/icarda/
	&ICRAF	Holding 190 species of wild, partially domesticated and domesticated trees	https://www.genebanks.org/genebanks/icraf/
	&ICRISAT	Holds more than 123,000 accessions of cultivated and wild relatives of pulses and cereals, including chickpea, sorghum and pigeonpea	https://www.genebanks.org/genebanks/icrisat/
	&ILRI	Collection of nearly 20,000 forage accessions, of which 97% are wild species	https://www.genebanks.org/genebanks/ilri/
	&IRRI	The largest collection of rice diversity in the world, with more than 130,000 accessions, including genetic stocks, landraces and wild relatives	https://www.genebanks.org/genebanks/irri/
	&IITA	Holding a collection of important African crops, including Bambara groundnut, cowpea, maize, soybean	https://www.genebanks.org/genebanks/iita/
Ex situ: Botanical gardens	BGCI	Including various databases such as PlantSearch, GardenSearch, ThreatSearch and GlobalTreeSearch	https://tools.bgci.org/garden_search.php
	BGCI-US	Botanic Gardens Conservation International and United States Botanic Garden collaboration effort identifying 22 global and 108 priority CWRs not reported in crop gene banks.	(Meyer and Barton., 2019)
In situ	ECPGR	Comprehensive concept for *in situ* conservation of CWR in Europe	https://www.ecpgr.cgiar.org/
	National Parks in India	Complete list of 103 national parks found in India currently under the protection of the Government	https://www.careerpower.in/national-parks-india.html
	RBGSYD	Australian *in situ* conservation site containing several key CWRs such as Macadamia nut, finger lime, etc	https://www.rbgsyd.nsw.gov.au/
	UNESCO	Biosphere reserves for the promotion of conserving biodiversity with sustainable use	https://en.unesco.org/biosphere
Software/ tools	GRIN-Global	Genebank information management system (open-sourced software)	(Postman et al., 2009) https://www.grin-global.org/
	CWPs	CWR phylogenetic classification system	(Viruel et al., 2021)
	plaBiPD	Phylogenetic relations among flowering plants with published genome sequences	https://www.plabipd.de/plant_genomes_pa.ep

(Continued)

TABLE 1 Continued

Type	Platform/ resource	Details	Reference/Access
	Mercator	plaBiPD associated protein annotation tool	https://www.plabipd.de/mercator_about.html
	Nordic	Provides towards the planning and implementation of active *in-situ* and *ex-situ* conservation of CWR at a national level	http://www.cropwildrelatives.org/conservation-toolkit/introduction/
	GBIF	The Global Biodiversity Information Facility is a global database for the distribution of crop wild relatives with over 5 million records (34% germplasm records, 66% herbarium records)	https://www.gbif.org/dataset/07044577-bd82-4089-9f3a-f4a9d2170b2e
	ECP-GR Natura 2000	A tool for protected area managers of Europe to help identify which CWR genera are likely to occur in the protected areas	https://www.ecpgr.cgiar.org/crop-wild-relatives-in-natura-2000
Genomic databases	PLAZA – Monocots	Access to whole genome sequencing data	https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v5_monocots/
	PLAZA - Dicots	Access to whole genome sequencing data	https://bioinformatics.psb.ugent.be/plaza/versions/plaza_v5_dicots/
	Germinate	A generic plant genetic database with several CWRs	https://germinateplatform.github.io/get-germinate/
	NCBI	General database for literature, genes, genomes and protein sequences of CWRs (Not-CWR specific)	https://www.ncbi.nlm.nih.gov /
	CerealsDB	SNP database for wheat	https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/indexNEW.php
	Brassica Information Portal (BIP)	Brassica specific	https://www.brassica.info/
	Genome Database for Rosaceae (GDR)	Rosaceae specific	https://www.rosaceae.org/
	GenRes Gateway	Access point to the European genetic resources for plants, forests and animals (including CWR)	https://www.genres.eu/
	ECPGR Central crop database	A database containing passport data, characteristics and primary evaluation data of the major collections of the respective crops.	https://www.ecpgr.cgiar.org/resources/germplasm- databases/ecpgr-central-crop-databases

International Center for Tropical Agriculture (CIAT), International Maize and Wheat Improvement Center (CIMMYT), International Potato Center (CIP), International Center for Agricultural Research in the Dry Areas (ICARDA), World Agroforestry (ICRAF), International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), International Livestock Research Institute (ILRI), International Rice Research Institute (IRRI) and International Institute of Tropical Agriculture (IITA), European Search Catalogue for Plant Genetic Resources (EURISCO) and the National Plant Germplasm System (NPGS), Botanic Gardens Conservation International (BGCI), Crop wild phylorelatives (CWPs), The International Center for Agricultural Research in the Dry Areas (ICARDA), NCBI – National Center for Biotechnology information.

pollinated breeding lines of traditional African vegetables, tomato, Capsicum pepper and soybean. The results show that introduced diversity through seed kits effectively improved local nutrition by facilitating access to various vegetables and also the introduction of new germplasm may slow down genetic erosion and enhance local vegetable diversity (Stoilova et al., 2019).

One of the main concerns across genebanks is the misclassification of species, as previously species identification was mostly based on morphological traits. However, recently the combination of traditional methods combined with molecular approaches, such as DNA barcoding, have improved the accuracy of species identification (van Bemmelen van der Plaat et al., 2021). For example, Mason et al., 2015 proved high-throughput genotyping approaches, such as a SNP array, is an effective methodology for species confirmation. They performed diversity assessment, using the Illumina Brassica 60K SNP array, across 180 Brassicaceae samples sourced from the Australian Grains Genebank and showed 76 of samples were misclassified (Mason et al., 2015). Through advances in genome sequencing technology and introduction of marker assisted breeding, the use of CWRs has intensified and with this growing interest it is important to keep information in the genebanks well documented and accurate. This is particularly important for use of CWRs in breeding programs where the success rate is highly dependent on the genetic distance between the species, particularly in approaches where crossing compatibility is important, it is crucial to have accurate information regarding the species taxonomy.

Botanical gardens are another *ex-situ* resource for germplasm; moreover they play a crucial role in preventing species extinction through integrated conservation actions (Mounce et al., 2017). Mounce et al. (2017) showed that botanic gardens contribute to the conservation of at least 105,634 species, representing 30% of all plant species diversity, including over 41% of known threatened species (Mounce et al., 2017). The Botanic Gardens Conservation International (BGCI) has the largest collection of living plants (Table 1). The

GardenSearch database, within BGCI, is the only global source for botanical gardens and includes information on over 3,755 botanical institutions worldwide. GardenSearch allows users to search botanical gardens based on their location (country) and their specific features or expertise (https://tools.bgci.org/garden_search.php). For example, based on information stored in GardenSearch, the botanical garden of South Australia has a collection of 40% of Australian flora including drought and salt tolerant plants. This information can facilitate the access and identification of plants with traits of interest for both breeding and research purposes. PlantSearch within the BGCI searches across 1,582,767 collection records, representing 642,718 taxa, at 1,194 institutions; in addition with Plant Search there is a specific option for CWR search at the taxa level (https://tools.bgci.org/plant_search.php) (Mounce et al., 2017).

In-situ resources

In contrast to *ex-situ* conservation sites, *in-situ* sites are typically natural habitats which are rarely curated, for example conservation/rehabilitation facilities or national parks. The benefit of *in-situ* resources is that they are genetically dynamic and continue to evolve in response to both natural and artificial selection, thereby enhancing their adaptation to the environments in which they are grown (Phillips et al., 2016). However, these *in-situ* collections are vulnerable to habitat destruction and/or encroachment caused by civil strife, human settlement pressure and natural disasters including wildfires, flooding, drought and volcanic eruptions. As such, the development of effective CWR conservation strategies is required nationally and globally. Several nations have already prioritised *in situ* CWR conservation, for example, Cyprus (178 priority CWR taxa) (Phillips et al., 2014), UK (148 priority CWR taxa) (Maxted et al., 2007; Fielder et al., 2015), US (821 priority CWR taxa) (Khoury et al., 2013; Khoury et al., 2019), Mexico (310 priority CWR taxa) (Contreras-Toledo et al, 2018), Czech (238 priority CWR taxa) (Taylor et al., 2013) and Norway (204 priority CWR taxa) (Phillips et al., 2016). These *in-situ* conservation efforts provide an ongoing roadmap for the study of the evolutionary history of the plant, which can provide insight into the persistence of traits, identification of new agriculturally significant traits and maintaining biodiversity (Khoury Colin et al., 2020). However, the incorporation of CWRs into traditional farming systems must be carefully considered as it may lead to unfavourable outcomes, for example, a study by Bernal et al., 2019., found that by incorporating a secluded maize genotype (Zea diploperennis) into Mexican and Argentinian farms, the pest 'corn leafhopper' was able to emerge as a widespread pest to corn farmers (Bernal et al., 2019).

Furthermore, CWR *in-situ* sites typically overlap with regions of high biodiversity, for example, as described by Vincent et al. (2022), the identified Mediterranean basin CWR hotspot shared 91% of its area with a region of high biodiversity, similarly, the California Floristic Province shared 90% between the CWR and biodiversity hotspots. This overlap has since been harnessed to aid in crop diversity and improvement studies, for example, the Unesco biosphere reserves promote solutions that reconcile the conservation of biodiversity with sustainable development (Benz et al., 2000). However, it is important to consider that *in-situ* resources should not only be limited to 'wild' regions. Traditional farming systems are not closed and isolated from gene flow, Louette et al., 1997., showed that the maize varieties cultivated by farmers of Cuzalapa, Mexico, changes in composition over time (Iltis et al., 1979; Louette et al., 1997). Despite certain changes to the germplasm being permanent, for example, the teosinte germplasm in maize which persists during advanced generations of backcrossing (Kato and Sanchez, 2002). In addition to the biodiversity hotspots, centers of origin/diversity, defined as global crop domestication regions including high diversity of both crops and their wild relatives (Vavilov, 1926), could be used as major sources for identification of CWRs. These diversity centers/regions include China; India; Indo-Malayan; Inner Asiatic; Mediterranean; Ethiopian; Central American; the Peruvian-Ecuadorian-Bolivian center, with sub-centers in both Chiloe, Chile and around the Brazil-Paraguay border (Vavilov et al., 1992; Pironon et al., 2020; Maxted and Vincent, 2021). Recently, by assessing the distribution of 222 major international crops and 2,731 of their wild relatives, including both closely and distant related wild species to the crops, Pironon et al. showed geographic distribution of major crop species and their closely related wild species strongly overlap with the Vavilov centers (Pironon et al., 2020). Identification of both crop and wild species diversity hotspots will provide opportunities for identifying and applying more focused conservation strategies for CWRs.

Considering CWRs have been neglected for years and there are many endangered species assessment of national and/or global *in-situ* resources to identify which CWRs are endangered or becoming extinct, whilst screening areas that are rich in wild crops and biodiversity (Hübner and Kantar, 2021) is crucial for protecting CWRs. For example, an assessment of wild banana species (*Musa* spp.) found that 11 out of 59 CWRs are vulnerable and another nine are endangered (Mertens et al., 2021). Khoury et al. (2019), found that of 600 CWR taxa assessed 7% may be critically endangered in their natural habitat and 50% may be endangered. These assessment programs involve a 'gap analysis' whereby the currently known and available CWR taxa (*in-situ/ex-situ* resources) are evaluated for their ability to provide future biodiversity to improve food security (Zair et al., 2021). By conducting a thorough gap analysis, Ng'uni et al., 2019., found that 459 CWR taxa out of a national Zambian inventory of 6305 taxa should now be included as part of their conservation and sustainability CWR checklist, with 59 to be specifically prioritised for future food security. The identified taxa represented an agriculturally significant group that was selected due to a shift in socio-economic values to ensure the nation's food security in the oncoming years. Several nations have conducted their own gap

analysis to ensure food security (Contreras-Toledo et al., 2019; Ng'uni et al., 2019; Tas et al., 2019; González-Orozco et al, 2021; Khaki Mponya et al., 2021; Rahman et al., 2021) and globally ten new *in-situ* conservation sites have been recommended as conservation zones to help achieve global food demand by expanding the *in-situ/ex-situ* resources (Zair et al., 2021).

To successfully establish *in-situ/ex-situ* resources to maintain and improve biodiversity, nations must create an inventory of all known plant taxa. These inventories provide a preliminary resource for the identification of critical taxa, such as CWRs (Teso et al., 2018; Allen et al., 2019; El Mokni et al., 2022). Whilst it is important for each nation to conduct an internal inventory, an unbiased global-scale inventory is also critical to establish CWR taxa. Vincent et al. (2013), originally created a global inventory of important CWR taxa, totaling 1667 taxa, divided between 37 families and 108 genera (Vincent et al., 2013). These inventories serve as the foundation for *in-situ/ex-situ* conservation, as they represent a 'living' CWR databank. However, as these taxa are truly wild, they will continue to evolve, and as such inventories only represent a snapshot of the population from the time of sampling, and recurring sampling is required to update inventories. A list of major global and national inventories is shown in Table 2.

Platforms: Tools for accessing, managing or utilising CWR data and metadata

Several platforms have begun to emerge with the explicit purpose of user-friendliness, designed to aid breeders and scientists alike (Raubach et al., 2021) to facilitate accessibility to CWR resources, including germplasm and genomic data (Table 1). These platforms attempt to solve the most common challenges in handling high throughput data from phenotyping to genotyping: 1) data format, 2) data sharing, 3) data versioning, and 4) historical data (Raubach et al., 2021). For example, GRIN-global (https://www.grin-global.org/) is open-source software for genebank workers to create and manage a genebank's data. Genesys and CGIAR are also examples of genebank platforms/databases (as discussed in the *ex-situ* section) that have been developed at a global scale to efficiently store and categorise data and facilitate the access and conservation of plant species including CWRs. Several other platforms are also available (discussed in the following sections) for visualizing, managing, accessing and storing large datasets related to crops and their relatives.

Software/tool-based platforms

Software/tool-based platforms are essential for data visualisation or organisation and help to gain a better understanding of the accessions stored in genebanks. For example, the Crop wild phylorelative platform (CWP in Table 1) (Viruel et al., 2021) helps to predict the phylogenetic distance (through housekeeping genes or whole genome analysis) and cytogenetic compatibility for breeding programs to help estimate the CWR gene pool classification (Brozynska et al., 2016; Viruel et al., 2021). Alternatively, plaBiPD provides an online platform that visualizes the phylogenetic relationship of genome sequences of flowering plants including CWRs. Furthermore, the associated Mercator online tool allows for the assignment of functional annotations to land plant protein sequences (Schwacke et al., 2019; Bolger et al., 2021).

Database management platforms

Database management tools provide a quick and easy to use platform for the access, management and use of data derived from breeding programs, research studies and trait identification programs using both CWRs and farmed crops. The genotyping platform Germinate v3 (Table 1) (Shaw et al., 2017; Raubach

TABLE 2 A list of major global and national CWR plant inventories.

Inventory name and details	Location assessed	Reference
Globally important CWR taxa	Global	(Vincent et al., 2013)
Important CWR taxa of Mexico	Mexico	(Contreras-Toledo et al, 2018)
National inventory of CWR in Spain	Spain	(Teso et al., 2018)
National inventories of CWR	Portugal	(Brehm et al., 2008)
CWR in USA	USA	(Khoury et al., 2013)
Enhancing and stating the UK CWR inventory	UK	(Fielder et al., 2015)
Prioritised CWR inventory of Italy	Italy	(Landucci et al., 2014)
Enhancing the CWR inventory of Scotland	Scotland	(Fielder et al., 2016)
Setting conservation priorities for CWR in the Fertile Crescent	Fertile Crescent	(Zair et al., 2018)
Prioritised inventory for Tunisia	Tunisia	(El Mokni et al., 2022)
CWR inventory of South, West and North Africa	South, West and North Africa	(Lala et al., 2018; Allen et al., 2019; Nduche et al., 2021)

et al., 2021) provides a rapid directory for importing and exporting plant genetic data such as erm plasm, markers, traits and locations. Germinate v3 has showcased its usefulness in breeding efforts that involve CWRs, specifically those associated with the Crop Trust Crop Wild Relatives project (https://www.cwrdiversity.org). Currently, Germinate v3 (20th of April, 2022) contains the directories for CWR taxa: Cowpea (~13100 germplasms), Finger Millet (~1600 germplasms), Grass Pea (~5600 germplasms), Pigeonpea (~2900 germplasms), Chickpea (~23500 germplasm), Alfalfa (~2700 germplasms), Carrot (248 germplasms), Pearl Millet (~2400 germplasms), Barley (~33200 germplasms), Wheat, Sorghum (~2800 germplasms), Eggplant (~3300 germplasms), Rice (~4900 germplasms) and Sunflower (~7900 germplasms) and DIIVA (~2900 germplasms). The use of Germinate has been employed in recent CWR studies. For example, Kouassi et al., 2021., generated interspecies hybrids with eggplants and nine related CWRs. The successfully generated hybrid lines were genotypically and phenotypically screened, wherein it was established that the drought tolerance traits were controlled by genes that are in linkage disequilibrium or have pleiotropic effects. The phenotypic characteristics have been stored in Germinate to provide access to both the user and breeders (Kouassi et al., 2021). Furthermore, Germinate also provides evaluation data of breeding programs. Metwally et al., 2021., generated 13 new superior F_{10} lines of cowpea by crossing CWRs, improving seed yield and seed quality, as well as introducing earlier maturation. The two datasets which cover 11 different traits for 15 cowpea accessions (total of 2640 data points) were uploaded to Germinate for visualization or downloads (Metwally et al., 2021).

Breeding and research resources are widely available for several crop species such as GrainGenes for wheat, barley, rye and oat (Blake et al., 2019), MaizeGDB for maize (Portwood et al., 2019) and SoyBase for soybean (Grant et al., 2010). These databases primarily host and facilitate the exploration of detailed breeding, pedigree, QTL and molecular information across crop populations. Whilst genomic information regarding CWRs may be presented in these databases, particularly in the case of family-wide databases such as the Sol Genomics Network for Solanaceae (Fernandez-Pozo et al., 2015), they are deposited with no tools for comparative analysis. The development of integrated tools accessible in comprehensive databases is needed to facilitate direct comparisons between wild and domesticated individuals.

Genomic databases

The PLAZA platform holds genomic data of both monocots and dicots. This platform compares the genomic data of submitted dicots and monocots to centralized genomic databases (Van Bel et al., 2022). The submitted genomic data is represented as an interactive phylogenetic tree style figure that links to a bioinformatic 'workbench'. The workbench includes tools such as gene family plots, collinearity statistic tools, localization tools and direct BLAST tools to the PLAZA protein sequences. Similarly to PLAZA, CerealsDB is a specific database platform for cereals like wheat (Wilkinson et al., 2020), providing several key features such as a SNP database for Axiom® 820K and 35K SNP arrays, KASP probes, iSelect Arrays, TaqMan® probes. The database is curated to provide agronomically important SNPs (e.g. flowering time associated markers). Furthermore, database platforms such as the Brassica information portal (Brassicaceae) (Eckes et al., 2017) and the Genome database for Rosaceae (Rosaceae) (Evans et al., 2013) have been established as a way to collate and exchange open source information relating to the Brassica and Rosaceae genomes and genetics, respectively, although the databases do not contain CWR resources directly, many of the projects included do include CWR resources. The Legume Information System and Legume Federation project provides an excellent collection of genomic and variant data for over 15 crop species, with a large range of accompanying CWR data (Dash et al., 2016).

Platform models that assist in data handling

A major issue in integrating informatics is a standardised model for data handling, especially as the information regarding the CWR conservation status and breeding programs is diverse and dispersed (Moore et al., 2008). These challenges can be identified by understanding the findable, accessible, interoperable and reusable (FAIR) curation and annotation of minor and underutilized crops (Andrés-Hernández et al., 2021). To address this, the European Crop Wild Diversity Assessment and Conservation Forum developed the Crop Wild Relative Information system (CWRIS) that incorporates an eXtensible Markup Language schema to aid data sharing and exchange. This system integrates with more partitions data into taxon-, site-, and population-specific elements, allowing for the integration with standard conservation biology (Kell et al., 2007; Kell et al., 2008; Moore et al., 2008). CWRIS was developed to provide access of the CWR data to a broader user community such as plant breeders, conservation and rehabilitation site managers, government, biologists and the wider public (Kell et al., 2007). CWRIS has since been integrated into GRIN-Global (https://npgsweb.ars-grin.gov/gringlobal/taxon/taxonomysearchcwr), as the website is no longer being maintained or updated.

Pangenomes to capture CWRs genetic variation

In recent years, advances in genome sequencing and bioinformatic tool development have extended the means to fully catalogue genetic variation among domestication and CWR

populations through the construction of pangenomes (Bayer et al., 2020; Jayakodi et al., 2021; Tay Fernandez et al., 2022). Pangenomes achieve this by providing a comprehensive genomic reference to which both small variants, including single-nucleotide polymorphisms (SNPs), and structural variants, including presence/absence variation of large nucleotide sections (PAVs), can be identified across diverse populations (Danilevicz et al., 2020). In addition, analysis of pangenomics allows for the more accurate predication of underlying genetics that are associated with phenotypic variation, such as transposable elements, recombination and double-stranded break/repair (Saxena et al., 2014; Dolatabadian et al., 2020; Song et al., 2020). As pangenomes excel in capturing large structural variation, as is increasingly found between highly divergent populations, they are ideally suited for the comparison of domesticated genomes to CWR taxa to capture 'wild genes' that would be overlooked when using a traditional reference genome (Khan et al., 2020). For example, a pangenome assembly of *Brassica oleracea* with 87 domesticated accessions (Bayer et al., 2021b) identified 58,347 genes across all individuals in comparison to a study that included 8 domesticated accessions and 1 CWR (Golicz et al, 2016) (8 landraces and 1 CWR), which identified a higher number of genes (63,865) (Golicz et al., 2016; Bayer et al., 2021b). Similar findings have been shown in sorghum (Tao et al., 2021) and rice (Xu et al., 2012), where the inclusion of CWR individuals led to large increases in the breadth of genes uncovered.

Beyond capturing more genes, the addition of CWR to pangenomes facilities the identification of novel SNPs and PAVs that are not found in domesticated populations. For example, Mace et al., 2021 performed comparative analysis in sorghum to quantify the 'contribution of CWR diversity' by establishing the average total number of SNPs per genotype. They found that wild/weedy species contained about one SNP every 763 bp compared to landraces that contained one SNP every 1,282 bp and inbred lines containing one SNP every 1,543 bp (Mace et al., 2021). Lam et al., 2010 also performed a comparative study between 17 wild and 14 cultivated soybean genomes showed higher diversity of SNPs and PAVs among wild species in compared to cultivated. In total, they found 6,318,109 SNPs and 186,177 PAVs, with the CWR genomes carrying 34.66% more SNPs (Lam et al., 2010). This is a clear indication that through optimising our agriculturally important crops, their respective genetic diversity has been reduced and CWR make promises to widen selection diversity (Nelson et al., 2018; Bailey-Serres et al., 2019).

Machine learning and CWRs

The application of machine learning (ML) has proven its efficiency in handling huge amounts of data and is becoming more popular in various plant science fields including gene identification and classification, and biodiversity analysis (Bayer et al., 2021a). For example, in Arabidopsis a ML model was developed to identify candidate stress-related genes by comparing whole genome expression data between the control and stress samples (Wegrzyn et al., 2014). In soybean, a ML model was developed to predict agronomically important traits, including yield, protein, oil, moisture and height, using SNP markers (Liu et al., 2019). Similarly, Ma et al., 2018., successfully developed a ML model to predict eight phenotypic traits among 2000 wheat individuals using 33,709 DArT (Diversity Array Technology) markers (Ma et al., 2018). ML is now also being used to predict mature yield in early development using a combination of image and genotype data (Danilevicz et al., 2021; Danilevicz et al., 2022). Recently ML models were developed for identification of core and dispensable genes in *Oryza sativa* L. and *Brachypodium distachyon* (L.) P. Beauv. using existing pangenomic information. The significant potential of these models is to identify core and dispensable genes in a new species without construction of pangenome (Yocca and Edger, 2022), such approaches can facilitate and speed up genes identification in new cultivated and wild species.

Understanding and usage of environmental conditions, in particular of CWR populations helps in selecting individual populations for the specific introgression goal. CWRs and landraces have occupied local niches (e.g., hot vs. cold regions) and have been shaped by natural selection (Cortés and López-Hernández, 2021), and these traits can be easily tracked when considering collection environmental site parameters. For example, Ariani et al, 2018, by using ~20,000 SNPs across 249 accession of wild *Phaseolus vulgaris*, identified 5 geographically distinct subpopulation, which mostly affected by temperature and rainfall of the regions (Ariani et al., 2018) Berny Mier Y. Teran et al., 2020, also documented that the lines driven from wild parents from the lower rainfall regions produced higher yield in both drought and watered conditions in compare to lines driven from domesticated parents (Berny Mier Y. Teran et al., 2020). Using ML algorithms is also a powerful approach to combine information of germplasm resources and environmental conditions for identification of candidate germplasms with traits of interest. This approach, finding adaptative traits based on environmental parameters, is known as FIGS (Focused Identification of Germplasm Strategy) (Khazaei et al., 2013). Several ML models based on the FIGS approach have been successfully developed and used for identifying germplasm of interest (Table 3). For instance, the identification and classification of *Vicia faba* genetic resources with traits related to drought tolerance (Khazaei et al., 2013). Similarly, in wheat, ML algorithms used for analysing accumulative stem rust trait data (1988-1994), and geographical data of accessions (including landraces and improved accessions) screened for stem rust over 2,000 collection sites revealed an association between the geographic distribution of resistance accessions and environmental variables at collection sites (Bari et al., 2012). Another ML model was successfully developed to predict stripe rust resistance in wheat, based on the stripe rust scores of 725 wheat landrace accessions

with collection site information associated with 2,910 accessions in the ICARDA genebank (Bari et al., 2014). Genetic diversity analysis among 80,000 wheat accessions (including 3,903 wild relatives) also revealed landraces with unexplored diversity and genetic footprints defined by regions under selection (Sansaloni et al., 2020). ML has facilitated the study and discovery of several genetic resources with agronomically valuable traits in crops. There are also "global database for the distribution of wild relatives" (https://www.gbif.org/dataset/07044577-bd82-4089-9f3a-f4a9d2170b2e) which includes the distribution data of crop wild relatives that can be used to extract geographical information and potential environmental conditions for CWRs.

Limitation to uses of CWRs within breeding programs

There are many challenges that still prevent the wide-spread use of CWRs as a source of superior alleles that can be incorporated into elite cultivated germplasm. The relatedness, compatibility and crossability of CWRs to their cultivated counterparts is one issue largely inhibiting the straightforward introduction of CWR traits through traditional breeding. For example, in cotton highly disease resistant sources were identified in wild diploid species, including *Gossypium. longicalyx* J.B. Hutch. & B.J.S. Lee; *G. somalense* (Gürke) J.B. Hutch.; *G. stocksii* Mast.; *G. arboreum* L.; and tetraploid species of *G. barbadense* L. (Yik and Birchfield, 1984); however due to genetic incompatibility, ploidy, climbing growth habit, photoperiodism, and agronomic issues breeders were unable to use these resources. Later, through the development of three-species hybrids, researchers were successfully able to introduce donor plants which were fertile and had reniform nematode resistance (Robinson et al., 2004; Konan et al., 2007).

Furthermore, trait identification and selection might be challenging and significantly affected by environment as there are radically different selection regimes in a wild state/region compared to a domesticated state/region while a trait can be useful in a domesticated state (and selected for) may not be useful in the wild and vice-versa. For example, Parker et al.,

TABLE 3 Case studies of the most recent applications of CWRs for crop improvement.

CWR	Application	Outcome	Reference
CWR of Cinnamomum, Piper, Vigna and Oryza in Sri Lanka	ML models to simulate the potential distribution across nine CWR species	The model was able to identify highly vulnerable species to climate change and predict the potential decrease in their suitable habitat by 2050. The study also identifies potential CWR rich areas for future *in-situ* conservation.	(Ratnayake et al., 2021)
ICARDA genebank barley accessions	FIGS *via* ML models	Providing predictive characterization for entire ICARDA barely collection	(Azough et al., 2019)
Wild blueberry	ML algorithms for yield prediction by evaluating bee species composition and weather factors	Prediction (with 93% accuracy) showed bee species composition and weather are significant in yield variability while wet rainy springs will greatly reduce blueberry yield.	(Obsie et al., 2020)
Wild cacao	Using ML model for surveying canopy and vegetation assessments	92% of classification accuracy for the structural attributes of the canopy	(Duarte-Carvajalino et al., 2021)
Large collection of *Vicia faba* L.	ML models used to evaluate FICS approach for identification of traits related to drought	The model was successful to indicate leaflet, canopy temperature and relative water content are important traits for drought-tolerance selection.	(Khazaei et al., 2013)
Solanum pimpinellifolium	Genome editing (*de novo* domestication	Produced a modified version of the wild *S. pimpinellifolium* which displayed a 10 times increase in the number of fruit and a 3 times increase in fruit size. The fruit also contained 500% more lycopene compared to the commonly cultivated *S. lucopersicum*.	(Ariani et al, 2018)
Physalis pruinose	Genome editing (*de novo* domestication	Edited orthologues of cultivated tomato in the distant relative *P. pruinose* to improve plant architecture, flower production and fruit size.	(Lemmon et al., 2018)
Oryza alta	Genome editing (*de novo* domestication	Established the first ever polyploid rice by genome editing the allotetraploid relative *O. alta*.	(Yu et al., 2021)
Aegilops tauschii	Association genetics with resistance gene enrichment sequencing (AgRenSeq)	Developed the AgRenSeq methodology and identified two novel wheat stem rust resistance genes, *Sr46* and *SrTA1662*, in a wild wheat progenitor.	(Arora et al., 2019)
Solanum americanum	Resistance gene enrichment sequencing and single-molecule real-time sequencing (SMRT RenSeq)	Identified the genome-wide repertoire of nucleotide-binding leucine-rich repeat type R genes in the wild *S. americanum* and cloned *Rpi-amr3i*, a novel R gene for potato late blight.	(Witek et al., 2016)
Oryza rufipogon	Genome editing	Optimised an efficient transformation system in wild rice, aiding future genome editing efforts including *de novo* domestication.	(Xiang et al., 2022)
Solanum peruvianum	Genome editing	Developed a genome editing approach using protoplast regeneration for the tetraploid wild tomato relative.	(Lin et al., 2022)

(2020), suggested the decreased-pod dehiscence (PD) trait among domesticated haplotypes of common bean is as a result of the different fitness landscape imposed by domestication, where stronger selection pressure were used against PD in arid condition of North Mexico compared to tropical lowlands (Andes), where environmental humidity masks susceptibility to PD and reducing selection pressure against it (Parker et al., 2020). It is also often challenging to accurately evaluate the yield of CWRs since they can display growth forms or traits that are difficult to manage, for example the wild progenitor of common bean has naturally dehiscent seed pods, making yield measurements arduous to obtain, and has a larger, less compact growth habit that is far less suitable for cultivated environments compared to cultivated common bean (Koinange et al., 1996). Even if beneficial wild derived traits are introgressed into elite material, they can often have a negative effect on yield or yield-related traits, through linkage drag. A common example is the introduction of biotic stress tolerance genes, for example disease resistance genes, which improve some resistance/tolerance but are detrimental to other agronomic traits (Brouwer and St Clair, 2004; Summers and Brown, 2013) Furthermore, after introducing genetic material from CWRs into an elite background, problems with sterility, often seen at the F_1 or BC_1 generation, can arise (Wang et al., 2020; Bohra et al., 2022).

There are also a number of challenges of CWR application in breeding that have been eased by availability of more genomic resources, and advances in laboratory techniques, as discussed in the following section. These include lack of information of gene-trait relationships in wild species, uncertainty of how allelic combinations will be expressed in different cultivated crop backgrounds and difficulties of transferring genes of interest into crops (Dempewolf et al., 2017).

Modern breeding and CWRs

There are now avenues to harness CWRs and overcome some of these barriers. For instance, wild-derived genes conferring desirable alleles can now be introduced through precise genome editing into elite backgrounds without the need for lengthy introgression regimes, bypassing the barriers of linkage drag and reduced fertility that so often complicate the use of CWRs (Bohra et al., 2021). These modern approaches, utilising the advances in genomics and genome editing, provide promising pathways to overcome long-standing challenges and push CWRs to the forefront of crop improvement. Table 3, included examples of successful application of CWRs for crop improvement *via* modern breeding approaches.

Genomics provides an avenue to explore the genetic diversity in CWRs and identify agronomically valuable genes or QTL. Sequencing CWRs followed by *de novo* assembly can generate reference assemblies that underpin downstream applications, such as the functional characterization of genes and targeted genome editing. Although initially lagging behind cultivated crop genomes, a number of CWRs assemblies are now becoming available, including relatives of barley, rice, soybean, tomato and wheat (Brozynska et al., 2016; Bohra et al., 2022). Often in combination with high-throughput phenotyping, these genome assemblies have enabled the identification of several important genes and QTL from CWRs, for example numerous disease resistance genes in wheat (Yahiaoui et al., 2009; Periyannan et al., 2013; Saintenac et al., 2013) and QTL associated with oil content in soybean (Zhou et al., 2015). High-quality assemblies based on third generation long read sequencing are now becoming the standard for reference genomes in major crops. Advances in long-read sequencing in terms of increased accessibility and lower price points, will be vital for the construction of high-quality long read assemblies in a broad range of CWRs, which will unlock an arsenal of beneficial CWR genetic diversity ready to be harnessed for crop improvement.

There are also recent genomic methodologies that have been developed to identify genes linked to specific traits; for instance resistance gene enrichment sequencing (RenSeq) is a methodology that targets, enriches and sequences *R* genes within any plant genome based on common *R* gene motifs (Jupe et al., 2013). To date, it has been used to capture nucleotide-binding-site leucine-rich repeat proteins (NLRs), receptor-like proteins (RLPs) and receptor-like kinases (RLKs), which represent the largest families of *R* genes (Jupe et al., 2013; Lin et al., 2020). Since its initial development, RenSeq has been combined with other approaches, including ethyl methanesulfonate (EMS) mutagenesis (MutRenSeq), single-molecule real-time sequencing (SMRT RenSeq) and association genetics (AgRenSeq). These combined workflows have rapidly identified and cloned causative *R* genes in a wild potato relative (Witek et al., 2016), wheat (Steuernagel et al., 2016), wild diploid wheat (Arora et al., 2019) and rye (Vendelbo et al., 2022). RenSeq is a promising alternative to whole genome sequencing for large scale *R* gene identification, and if utilised in CWRs, has the potential to rapidly expand the *R* gene arsenal used for breeding disease resistant cultivars. Notably, AgRenSeq does not rely on a reference genome (Arora et al., 2019), therefore it is extremely applicable to CWRs that are yet to have a reference assembly, but whose cultivated counterpart has well characterised *R* genes.

While there has been rapid progress within the field of plant genome editing, the application within CWRs has been far slower. The limited genomic resources for many CWRs serves as an initial barrier, then the lack of functionally characterized gene targets and easy delivery system for those targets proves arduous. In spite of these challenges, one innovative application of CRISPR recently proposed is the manipulation of genes controlling important agronomic traits, for example plant architecture genes, while purposefully retaining valuable wild-

derived traits such as stress tolerance or improved nutritional quality; in essence, the domestication of a CWR or landrace that has never been cultivated. This approach, termed *de novo* domestication, can produce new crops from a CWR in a matter of generations through genome editing technology (Gasparini et al., 2021). Using a wild tomato relative, Zsögön et al., 2018., edited four key tomato domestication genes, *SELF-PRUNING*, *OVATE*, *FRUIT WEIGHT 2.2* and *LYCOPENE BETACYCLASE*, to produce an engineered tomato crop boasting increased fruit number and size compared to the wild parent, and vastly improved nutritional quality compared to cultivated tomato (Zsögön et al., 2018). A similar approach was undertaken in the orphan crop groundcherry, a distant tomato relative, whereby productivity traits including plant architecture, flower production and fruit size were improved by editing known tomato orthologues with CRISPR-Cas9 (Lemmon et al., 2018). One ambitious study utilised *de novo* domestication to develop the first ever polyploid rice crop, through the rapid domestication of an allotetraploid wild rice, *Oryza alta* (Yu et al., 2021). This has demonstrated a feasible route to create polyploid versions of diploid crops, which are said to benefit from genome buffering *via* gene redundancy, hybrid vigour and environmental fortitude (Mason and Batley, 2015). As researchers characterise more genes related to key domestication traits in model or major crops and high-quality CWR genome assemblies are generated, the potential for editing these genes in CWRs skyrockets, leading to the possible creation of new crops through *de novo* domestication. Furthermore, simultaneously identifying and cataloguing agronomically beneficial traits in CWRs will greatly enhance our ability to exploit wild genetic diversity, meaning *de novo* domesticated crops will be more nutritious and climate resilient than their cultivated relatives.

Despite the promising potential of *de novo* domestication, one of the major challenges preventing the widespread deployment of CRISPR in CWRs, and therefore *de novo* domestication, is the delivery system of the genome editing reagents. Even for elite cultivars, quick and easy methods for delivery that are widely transferable between species remain elusive (Zhan et al., 2021). The most popular DNA delivery approaches include agrobacterium-mediated delivery, which utilises the soil pathogen *Agrobacterium tumefaciens* to transfer DNA into the host genome, and biolistic or micro-projectile-mediated delivery, where the donor DNA is mechanically forced into the host cells (Ran et al., 2017). However, these methods come with certain limitations. *Agrobacterium*-mediated delivery is hindered by its inability to introduce small donor fragments, its difficulty in preventing plasmid integration and thereby producing a transgenic plant, and is dependent on the genotype of the recipient, particularly for monocot plants (Ran et al., 2017). While biolistic methods provide some advantages over *Agrobacterium*-mediated delivery, for example the delivery of multiple targets, its use is lower than expected due to issues with multiple copies of the transgene being incorporated into the host, resulting in altered gene expression or complete silencing. Efficient delivery methods using these approaches, after significant optimisation, have been established in model plants and select major crops. However, such methods are not easily transferrable to CWRs, as they often represent a diverse set of morphotypes which introduces unique challenges hindering delivery. On top of this, CWRs are also difficult to regenerate, further complicating the transformation process (Zhu et al., 2020).

Several alternative approaches for reagent delivery which were initially developed in animal cells, are being explored in plants (Ghogare et al., 2021). For example, a biolistics approach using nanoparticles offers a less harmful delivery method compared to larger microparticles, which may reduce delivery damage, a common issue encountered in plants due to the presence of a cell wall (Zhang et al., 2019; Cunningham et al., 2020). Most excitingly, delivery mediated by viral vectors can completely bypass the need for regeneration which is an extremely promising prospect for editing hard to regenerate CWRs, however this method is limited by its delivery capacity (Shan-e-Ali Zaidi and Mansoor, 2017). Novel delivery methods will help to overcome the barriers preventing widespread plant transformation and reduce the amount of optimisation needed. In doing so, efficient genome editing in CWRs will be one step closer.

Another potential approach for CWRs utilization in breeding schemes is through speed breeding. The concept of speed breeding revolves around manipulating the photoperiod (e.g. 12 hr extended to 22 hr) and temperature in a controlled growth facility to rapidly produce multiple crop generations per year (Watson et al., 2018). Through speed breeding, the genetic background of cultivars can be fixed in an accelerated timeframe, a process which usually takes years of inbreeding. Speed breeding has been tested and effectively produced multiple generations in a single year for crops such as barley, canola, chickpea, pea, rice, sorghum and wheat (Espósito et al., 2012; Rizal et al., 2014; Watson et al., 2018; Nagatoshi and Fujita, 2019; Rana et al., 2019). In the absence of precise genome editing, desirable traits from CWRs which are introgressed into elite cultivars through traditional breeding will often bring with them unwanted deleterious alleles. Hence, speed breeding can facilitate the quick growth of multiple generations, allowing undesirable traits to be selected against, and for these new varieties to reach a stable genetic background. In addition, speed breeding would benefit alternative approaches to domesticate CWRs without the use of CRISPR, such as germplasm conversion (Stephens et al., 1967; Rosenow et al., 1997; Klein et al., 2016). Germplasm conversion involves the alteration of germplasm through crossing, multiple rounds of selection for various traits and inbreeding to become well-adapted to new environments while also having favourable agronomic traits (Stephens et al., 1967). Extensive germplasm conversion has been done in Sorghum to transform numerous exotic varieties into early-maturing and dwarf-height varieties

that are adapted for cultivation in the US or other temperate regions (Stephens et al., 1967; Rosenow et al., 1997; Klein et al., 2016). As an alternative to genome editing, germplasm conversion could be harnessed to introduce important agronomic traits into CWRs through hybridization and then followed by marker-assisted selection (MAS). The advantage of this over genome editing is that specific knowledge of the target sequences is not required, only knowledge of the genomic region conferring the domestication trait/s. However, it is likely that this method would be more laborious and time consuming compared to genome editing approaches, as several generations are usually required to achieve the final product. Therefore, exposing these CWRs to speed breeding conditions may help to mitigate the time required for cycling multiple generations that is necessary for effective germplasm conversion of CWRs into commercially viable crops (Bhatta et al., 2021).

Conclusion

Crop wild relatives have remained under-utilised during crop domestication and intense crop breeding, despite the fact they harbour beneficial traits such as disease and pest resistance, and tolerance to abiotic stresses. CWRs have the potential to widen selection sources for breeders beyond the existing variation among cultivated crops to meet future foods' quality and quantity demands. A multi-resource integrative approach that utilises many of the resources outlined here will enable CWRs to be effectively used as a source of valuable genetic diversity. For example, ML strategies based on FIGS in combination with genomic and pangenomic resources that capture the gene diversity that exists in CWRs, will help to rapidly identify adaptative traits based on environmental parameters which will in turn guide the identification of genes underpinning these traits. However, realisation and utilisation of the full potential of the genes and diversity presented in CWRs will ultimately depend on the availability of resources and experimental techniques to support breeding programs (Hajjar and Hodgkin, 2007). There are a number of resources and databases that both researchers and breeders can benefit from, but ongoing efforts are crucial to keep these data well organised and up-to-date. This is only possible with the great collaboration between ecological/biological conservation sectors, who manage CWR ex/in -situ conservation and prevent extinction, researchers in the field of computer science, plant biology, for example plant genomics and agricultural industries, who assist with identification of traits/genes of interest among CWRs and only with this multidisciplinary effort is there a chance to guarantee the future food demands.

Author contributions

ST and JB conceptualized the review. ST wrote the main text with additions from WT, JZ, JM, DE and JB. DE and JB edited the paper. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by the Australian Research Council projects DP200100762, DP210100296 and the Grains Research and Development Corporation (UWA1905-006RTX).

Acknowledgments

WT would like to acknowledge the support of the Grains Research and Development Corporation.

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

Abdallah, F., Kumar, S., Amri, A., Mentag, R., Kehel, Z., Mejri, R. K., et al. (2021). Wild lathyrus species as a great source of resistance for introgression into cultivated grass pea (Lathyrus sativus l.) against broomrape weeds (Orobanche crenata forsk. and orobanche foetida poir.). *Crop Sci.* 61 (1), 263–276. doi: 10.1002/csc2.20399

Aday, S., and Aday, M. S. (2020). Impact of COVID-19 on the food supply chain. *Food Qual. Saf.* 4 (4), 167–180. doi: 10.1093/fqsafe/fyaa024

Allaby, R. G., Ware, R. L., and Kistler, L. (2019). A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evol. Appl.* 12 (1), 29–37. doi: 10.1111/eva.12680

Allen, E., Gaisberger, H., Brehm, J. M., Maxted, N., Thormann, I., Lupupa, T., et al. (2019). A crop wild relative inventory for southern Africa: A first step in linking conservation and use of valuable wild populations for enhancing food security. *Plant Genet. Resour.* 17 (2), 128–139. doi: 10.1017/S1479262118000515

Andrés-Hernández, L., Halimi, R. A., Mauleon, R., Mayes, S., Baten, A., and King, G. J. (2021). Challenges for FAIR-compliant description and comparison of crop phenotype data with standardized controlled vocabularies. *Database* 2021, 1–11. doi: 10.1093/database/baab028

Ariani, A., Berny Mier, Y. T. J. C., and Gepts, P. (2018). Spatial and temporal scales of range expansion in wild phaseolus vulgaris. *Mol. Biol. Evol.* 35 (1), 119–131. doi: 10.1093/molbev/msx273

Arora, S., Steuernagel, B., Gaurav, K., Chandramohan, S., Long, Y., Matny, O., et al. (2019). Resistance gene cloning from a wild crop relative by sequence capture and association genetics. *Nat. Biotechnol.* 37 (2), 139–143. doi: 10.1038/s41587-018-0007-9

Azough, Z., Kehel, Z., Benomar, A., Bellafkih, M., and Amri, A. (2019). "Predictive characterization of ICARDA genebank barley accessions using FIGS and machine learning," in *Intelligent environments (Workshops)*, 121–129.

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D., and Schroeder, J. I. (2019). Genetic strategies for improving crop yields. *Nature* 575 (7781), 109–118. doi: 10.1038/s41586-019-1679-0

Bari, A., Amri, A., Street, K., Mackay, M., De Pauw, E., Sanders, R., et al. (2014). Predicting resistance to stripe (yellow) rust (Puccinia striiformis) in wheat genetic resources using focused identification of germplasm strategy. *J. Agric. Sci.* 152 (6), 906–916. doi: 10.1017/S0021859613000543

Bari, A., Street, K., Mackay, M., Endresen, D. T. F., De Pauw, E., and Amri, A. (2012). Focused identification of germplasm strategy (FIGS) detects wheat stem rust resistance linked to environmental variables. *Genet. Resour. Crop Evol.* 59 (7), 1465–1481. doi: 10.1007/s10722-011-9775-5

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6 (8), 914–920. doi: 10.1038/s41477-020-0733-0

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., and Edwards, D. (2021a). The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome* 14 (3), e20112. doi: 10.1002/tpg2.20112

Bayer, P. E., Scheben, A., Golicz, A. A., Yuan, Y., Faure, S. , Lee, H., et al. (2021b). Modelling of gene loss propensity in the pangenomes of three brassica species suggests different mechanisms between polyploids and diploids. *Plant Biotechnol. J.* 19 (12), 2488–2500. doi: 10.1111/pbi.13674

Bayer, P. E., Valliyodan, B., Hu, H., Marsh, J. I., Yuan, Y., Vuong, T. D., et al. (2022). Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding. *Plant Genome* 15 (1), e20109. doi: 10.1002/tpg2.20109

Benz, B. F., Cevallos E, J., Santana M, F., Rosales A, J., and Graf, M. „ S. (2000). Losing knowledge about plant use in the sierra de manantlan biosphere reserve, Mexico. *Economic Bot.* 54 (2), 183–191. doi: 10.1007/BF02907821

Bernal, J. S., Dávila-Flores, A. M., Medina, R. F., Chen, Y. H., Harrison, K. E., and Berrier, K. A. (2019). Did maize domestication and early spread mediate the population genetics of corn leafhopper? *Insect Sci.* 26 (3), 569–586. doi: 10.1111/1744-7917.12555

Berny Mier Y. Teran, J. C., Konzen, E. R., Palkovic, A., Tsai, S. M., and Gepts, P. (2020). Exploration of the yield potential of mesoamerican wild common beans from contrasting eco-geographic regions by nested recombinant inbred populations. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00346

Bhatta, M., Sandro, P., Smith, M. R., Delaney, O., Voss-Fels, K. P., Gutierrez, L., Hickey, L. T., et al. (2021). Need for speed: Manipulating plant growth to accelerate breeding cycles. *Current Opin. in Plant Biol.* 60, 101986. doi: 10.1016/j.pbi.2020.101986

Blake, V. C., Woodhouse, M. R., Lazo, G. R., Odell, S. G., Wight, C. P., Tinker, N. A., et al. (2019). GrainGenes: centralized small grain resources and digital platform for geneticists and breeders. *Database (Oxford)* 2019, 1–7. doi: 10.1093/database/baz065

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S. R., et al. (2021). Reap the crop wild relatives for breeding future crops. *Trends Biotechnol.*

Bohra, A., Kilian, B., Sivasankar, S., Caccamo, M., Mba, C., McCouch, S. R., et al. (2022). Reap the crop wild relatives for breeding future crops. *Trends Biotechnol.* 40 (4), 412–431. doi: 10.1016/j.tibtech.2021.08.009

Bolger, M., Schwacke, R., and Usadel, B. (2021). "MapMan visualization of RNA-seq data using Mercator4 functional annotations," in *Solanum tuberosum* (New York, NY: Humana), 195–212.

Brehm, J. M., Maxted, N., Ford-Lloyd, B. V., and Martins-Louçao, M. A. (2008). National inventories of crop wild relatives and wild harvested plants: case-study for Portugal. *Genet. Resour. Crop Evol.* 55 (6), 779–796. doi: 10.1007/s10722-007-9283-9

Brouwer, D. J., and St Clair, D. A. (2004). Fine mapping of three quantitative trait loci for late blight resistance in tomato using near isogenic lines (NILs) and sub-NILs. *Theor. Appl. Genet.* 108 (4), 628–638. doi: 10.1007/s00122-003-1469-8

Brozynska, M., Furtado, A., and Henry, R. J. (2016). Genomics of crop wild relatives: expanding the gene pool for crop improvement. *Plant Biotechnol. J.* 14 (4), 1070–1085. doi: 10.1111/pbi.12454

Castañeda-Álvarez, N. P., Khoury, C. K., Achicanoy, H. A., Bernau, V., Dempewolf, H., Eastwood, R. J., et al. (2016). Global conservation priorities for crop wild relatives. *Nat. Plants* 2 (4), 16022. doi: 10.1038/nplants.2016.22

Contreras-Toledo, A. R., Cortés-Cruz, M. A., Costich, D., de Lourdes Rico-Arce, M., Brehm, J. M., and Maxted, N. (2018). A crop wild relative inventory for Mexico. *Crop Sci.* 58 (3), 1292–1305. doi: 10.2135/cropsci2017.07.0452

Contreras-Toledo, A. R., Cortés-Cruz, M., Costich, D. E., de Lourdes Rico-Arce, M., Brehm, J. M., and Maxted, N. (2019). Diversity and conservation priorities of crop wild relatives in Mexico. *Plant Genet. Resour.* 17 (2), 140–150. doi: 10.1017/S1479262118000540

Cortés, A. J., and López-Hernández, F. (2021). Harnessing crop wild diversity for climate change adaptation. *Genes* 12 (5). doi: 10.3390/genes12050783

Cunningham, F. J., Demirer, G. S., Goh, N. S., Zhang, H., and Landry, M. P. (2020). "Nanobiolistics: An emerging genetic transformation approach," in *Biolistic DNA delivery in plants* (New York, NY: Humana), 141–159.

Danilevicz, M. F., Bayer, P. E., Boussaid, F., Bennamoun, M., and Edwards, D. (2021). Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sens.* 13 (19), 3976. doi: 10.3390/rs13193976

Danilevicz, M. F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P. E., et al. (2022). Plant genotype to phenotype prediction using machine learning. *Front. Genet.* 13. doi: 10.3389/fgene.2022.822173

Danilevicz, M. F., Tay Fernandez, C. G., Marsh, J. I., Bayer, P. E., and Edwards, D. (2020). Plant pangenomics: approaches, applications and advancements. *Curr. Opin. Plant Biol.* 54, 18–25. doi: 10.1016/j.pbi.2019.12.005

Dash, S., Campbell, J. D., Cannon, E. K. S., Cleary, A. M., Huang, W., Kalberer, S. R., et al. (2016). Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Res.* 44 (D1), D1181–D1188. doi: 10.1093/nar/gkv1159

Dempewolf, H., Baute, G., Anderson, J., Kilian, B., Smith, C., and Guarino, L. (2017). Past and future use of wild relatives in crop breeding. *Crop Sci.* 57 (3), 1070–1082. doi: 10.2135/cropsci2016.10.0885

Dida, M. M., Oduori, C. A., Manthi, S. J., Avosa, M. O., Mikwa, E. O., Ojulong, H. F., et al. (2021). Novel sources of resistance to blast disease in finger millet. *Crop Sci.* 61 (1), 250–262. doi: 10.1002/csc2.20378

Dolatabadian, A., Bayer, P. E., Tirnaz, S., Hurgobin, B., Edwards, D., and Batley, J. (2020). Characterization of disease resistance genes in the brassica napus pangenome reveals significant structural variation. *Plant Biotechnol. J.* 18 (4), 969–982. doi: 10.1111/pbi.13262

Duarte-Carvajalino, J. M., Paramo-Alvarez, M., Ramos-Calderón, P. F., and González-Orozco, C. E. (2021). Estimation of canopy attributes of wild cacao trees using digital cover photography and machine learning algorithms. *iForest - Biogeosciences Forestry* 14 (6), 517–521. doi: 10.3832/ifor3936-014

Eckes, A. H., Gubała, T., Nowakowski, P., Szymczyszyn, T., Wells, R., Irwin, J. A., et al. (2017). Introducing the brassica information portal: Towards integrating genotypic and phenotypic brassica crop data. *F1000Research* 6:465. doi: 10.12688/f1000research.11301.1

El Mokni, R., Barone, G., Maxted, N., Kell, S., and Domina, G. (2022). A prioritised inventory of crop wild relatives and wild harvested plants of Tunisia. *Genet. Resour. Crop Evol.* 1–34, 1787–1816. doi: 10.1079/9781845930998.0471

Espósito, M., Almirón, P., Gatti, I., Cravero, V. P., Anido, F. S. L., and Cointry, E. (2012). A rapid method to increase the number of F1 plants in pea (*Pisum sativum*) breeding programs. *Genet. Mol. Res.* 11 (3), 2729–2732. doi: 10.4238/2012.June.18.1

Esquinas-Alcázar, J. (2005). Protecting crop genetic diversity for food security: political, ethical and technical challenges. *Nat. Rev. Genet.* 6 (12), 946. doi: 10.1038/nrg1729

Evans, K., Jung, S., Lee, T., Brutcher, L., Cho, I., Peace, C., et al. (2013). Addition of a breeding database in the genome database for rosaceae. *Database* 2013. doi: 10.1093/database/bat078

FAO (1999) *What is happening to agrobiodiversity?* Available at: https://www.fao.org/3/y5609e/y5609e02.htm.

Fernandez-Pozo, N., Menda, N., Edwards, J. D., Saha, S., Tecle, I. Y., Strickler, S. R., et al. (2015). The sol genomics network (SGN)–from genotype to phenotype to breeding. *Nucleic Acids Res.* 43 (Database issue), D1036–D1041. doi: 10.1093/nar/gku1195

Fielder, H., Brotherton, P., Hosking, J., Hopkins, J. J., Ford-Lloyd, B., and Maxted, N. (2015). Enhancing the conservation of crop wild relatives in England. *PloS One* 10 (6), e0130804. doi: 10.1371/journal.pone.0130804

Fielder, H., Smith, C., Ford-Lloyd, B., and Maxted, N. (2016). Enhancing the conservation of crop wild relatives in Scotland. *J. Nat. Conserv.* 29, 51–61. doi: 10.1016/j.jnc.2015.11.002

Gaikwad, K. B., Rani, S., Kumar, M., Gupta, V., Babu, P. H., Bainsla, N. K., et al. (2020). Enhancing the nutritional quality of major food crops through

conventional and genomics-assisted breeding. *Front. Nutr.* 7, 533453. doi: 10.3389/fnut.2020.533453

Gasparini, K., Moreira, J. D. R., Peres, L. E. P., and Zsögön, A. (2021). *De novo* domestication of wild species to create crops with increased resilience and nutritional value. *Curr. Opin. Plant Biol.* 60, 102006–102006. doi: 10.1016/j.Pbi.2021.102006

Gepts, P., Osborn, T. C., Rashka, K., and Bliss, F. A. (1986). Phaseolin-protein variability in wild forms and landraces of the common Bean(Phaseolus vulgaris): Evidence for multiple centers of domestication. *Economic Bot.* 40 (4), 451–468. doi: 10.1007/BF02859659

Ghogare, R., Ludwig, Y., Bueno, G. M., Slamet-Loedin, I. H., and Dhingra, A. (2021). Genome editing reagent delivery in plants. *Transgenic Res.* 30, 321–335. doi: 10.1007/s11248-021-00239-w

Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016). The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7, 13390. doi: 10.1038/ncomms13390

González-Orozco, C. E., Sosa, C. C., Thornhill, A. H., and Laffan, S. W. (2021). Phylogenetic diversity and conservation of crop wild relatives in Colombia. *Evolutionary Appl.* 14 (11), 2603–2617. doi: 10.1111/eva.13295

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38 (suppl_1), D843–D846. doi: 10.1093/nar/gkp798

Gregory, P. J., Johnson, S. N., Newton, A. C., and Ingram, J. S. (2009). Integrating pests and pathogens into the climate change/food security debate. *J. Exp. Bot.* 60 (10), 2827–2838. doi: 10.1093/jxb/erp080

Hajjar, R., and Hodgkin, T. (2007). The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* 156 (1), 1–13. doi: 10.1007/s10681-007-9363-0

Hübner, S., and Kantar, M. B. (2021). Tapping diversity from the wild: From sampling to implementation. *Front. Plant Sci.* 12 (38). doi: 10.3389/fpls.2021.626565

Iltis, H. H., Doebley, J. F., Guzmán M, R., and Pazy, B. (1979). Zea diploperennis (Gramineae): A new teosinte from Mexico. *Science* 203 (4376), 186–188. doi: 10.1126/science.203.4376.186

Inagaki, N., Asami, H., Hirabayashi, H., Uchino, A., Imaizumi, T., and Ishimaru, K. (2021). A rice ancestral genetic resource conferring ideal plant shapes for vegetative growth and weed suppression. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.748531

IPCC (2014). "Climate change 2014: Synthesis report," in *Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change*. Eds. R. K. Pachauri and L. A. Meyer (Geneva, Switzerland: IPCC).

Jayakodi, M., Schreiber, M., Stein, N., and Mascher, M. (2021). Building pan-genome infrastructures for crop plants and their use in association genetics. *DNA Res.* 28 (1), dsaa030. doi: 10.1093/dnares/dsaa030

Jupe, F., Witek, K., Verweij, W., Śliwka, J., Pritchard, L., Etherington, G. J., et al. (2013). Resistance gene enrichment sequencing (RenSeq) enables reannotation of the NB-LRR gene family from sequenced plant genomes and rapid mapping of resistance loci in segregating populations. *Plant J.* 76 (3), 530–544. doi: 10.1111/tpj.12307

Kato, T., and Sanchez, J. (2002). Introgression of chromosome knobs from zea diploperennis into maize [Zea mays l.]. *Maydica (Italy)* 47(1), 33–5.

Kell, S., Jury, S., Knüpffer, H., Ford-Lloyd, B., and Maxted, N. (2007). PGR forum: serving the crop wild relative user community. *Bocconea* 21, 413–421.

Kell, S., Moore, J., Iriondo, J., Scholten, M., Ford-Lloyd, B., and Maxted, N. (2008). CWRIS: an information management system to aid crop wild relative conservation and sustainable use. *Crop wild relative conservation and use* (Wallingford UK: CABI), 471–491. doi: 10.1079/9781845930998.047

Khaki Mponya, N., Chanyenga, T., Magos Brehm, J., and Maxted, N. (2021). *In situ* and ex situ conservation gap analyses of crop wild relatives from Malawi. *Genet. Resour. Crop Evol.* 68 (2), 759–771. doi: 10.1007/s10722-020-01021-3

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., and Varshney, R. K. (2020). Super-pangenome by integrating the wild side of a species for accelerated crop improvement. *Trends Plant Sci.* 25 (2), 148–158. doi: 10.1016/j.tplants.2019.10.012

Khazaei, H., Street, K., Bari, A., Mackay, M., and Stoddard, F. L. (2013). The FIGS (Focused identification of germplasm strategy) approach identifies traits related to drought adaptation in vicia faba genetic resources. *PloS One* 8 (5), e63107. doi: 10.1371/journal.pone.0063107

Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M. M., et al. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytol.* 233 (1), 84–118. doi: 10.1111/nph.17733

Khoury Colin, K., Carver, D., Greene Stephanie, L., Williams Karen, A., Achicanoy Harold, A., Schori, M., et al. (2020). Crop wild relatives of the united states require urgent conservation action. *Proc. Natl. Acad. Sci.* 117 (52), 33351–33357. doi: 10.1073/pnas.2007029117

Khoury, C. K., Greene, S. L., Krishnan, S., Miller, A. J., and Moreau, T. (2019). A road map for conservation, use, and public engagement around north america's crop wild relatives and wild utilized plants. *Crop Sci.* 59 (6), 2302–2307. doi: 10.2135/cropsci2019.05.0309

Khoury, C. K., Greene, S., Wiersema, J., Maxted, N., Jarvis, A., and Struik, P. C. (2013). An inventory of crop wild relatives of the united states. *Crop Sci.* 53 (4), 1496–1508. doi: 10.2135/cropsci2012.10.0585

Klein, R. R., Miller, F. R., Bean, S., and Klein, P. E. (2016). Registration of 40 converted germplasm sources from the reinstated sorghum conversion program. *J. Plant Registrations* 10 (1), 57–61. doi: 10.3198/jpr2015.05.0034crg

Koinange, E. M., Singh, S. P., and Gepts, P. (1996). Genetic control of the domestication syndrome in common bean. *Crop Sci.* 36 (4), 1037–1045. doi: 10.2135/cropsci1996.0011183X003600040037x

Konan, O. N., D'Hont, A., Baudoin, J. P., and Mergeai, G. (2007). Cytogenetics of a new trispecies hybrid in cotton: [(Gossypium hirsutum l. × g. thurberi Tod.)2 × g. longicalyx hutch. & Lee]. *Plant Breed.* 126 (2), 176–181. doi: 10.1111/j.1439-0523.2007.01325.x

Kouassi, A. B., Kouassi, K. B. A., Sylla, Z., Plazas, M., Fonseka, R. M., Kouassi, A., et al. (2021). Genetic parameters of drought tolerance for agromorphological traits in eggplant, wild relatives, and interspecific hybrids. *Crop Sci.* 61 (1), 55–68. doi: 10.1002/csc2.20250

Lala, S., Amri, A., and Maxted, N. (2018). Towards the conservation of crop wild relative diversity in north Africa: checklist, prioritisation and inventory. *Genet. Resour. Crop Evol.* 65 (1), 113–124. doi: 10.1007/s10722-017-0513-5

Lam, H.-M., Xu, X., Liu, X., Chen, W., Yang, G., Wong, F.-L., et al. (2010). Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat. Genet.* 42 (12), 1053–1059. doi: 10.1038/ng.715

Landucci, F., Panella, L., Lucarini, D., Gigante, D., Donnini, D., Kell, S., et al. (2014). A prioritized inventory of crop wild relatives and wild harvested plants of Italy. *Crop Sci.* 54 (4), 1628–1644. doi: 10.2135/cropsci2013.05.0355

Lemmon, Z. H., Reem, N. T., Dalrymple, J., Soyk, S., Swartwood, K. E., Rodriguez-Leal, D., et al. (2018). Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* 4, 766–770. doi: 10.1038/s41477-018-0259-x

Lin, X., Armstrong, M., Baker, K., Wouters, D., Visser, R. G. F., Wolters, P. J., et al. (2020). *RLP/K* enrichment sequencing; a novel method to identify receptor-like protein (*RLP*) and receptor-like kinase (*RLK*) genes. *New Phytol.* 277 (4), 1264–1276. doi: 10.1111/nph.16608

Lin, C.-S., Hsu, C.-T., Yuan, Y.-H., Zheng, P.-X., Wu, F.-H., Cheng, Q.-W., et al. (2022). DNA-Free CRISPR-Cas9 gene editing of wild tetraploid tomato solanum peruvianum using protoplast regeneration. *Plant Physiol.* 188 (4), 1917–1930. doi: 10.1093/plphys/kiac022

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-genome of wild and cultivated soybeans. *Cell* 182 (1), 162–176.e113. doi: 10.1016/j.cell.2020.05.023

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., and Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. *Front. Genet.* 10. doi: 10.3389/fgene.2019.01091

Louette, D., Charrier, A., and Berthaud, J. (1997). *In situ* conservation of maize in Mexico: Genetic diversity and maize seed management in a traditional community. *Economic Bot.* 51 (1), 20–38. doi: 10.1007/BF02910401

Mace, E. S., Cruickshank, A. W., Tao, Y., Hunt, C. H., and Jordan, D. R. (2021). A global resource for exploring and exploiting genetic variation in sorghum crop wild relatives. *Crop Sci.* 61 (1), 150–162. doi: 10.1002/csc2.20332

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., et al. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. *Planta* 248 (5), 1307–1318. doi: 10.1007/s00425-018-2976-9

Mason, A. S., and Batley, J. (2015). Creating new interspecific hybrid and polyploid crops. *Trends Biotechnol.* 33 (8), 436–441. doi: 10.1016/j.Tibtech.2015.06.004

Mason, A. S., Zhang, J., Tollenaere, R., Vasquez Teuber, P., Dalton-Morgan, J., Hu, L., et al. (2015). High-throughput genotyping for species identification and diversity assessment in germplasm collections. *Mol. Ecol. Resour* 15 (5), 1091–1101. doi: 10.1111/1755-0998.12379

Maxted, N. (2008). *Crop wild relative conservation and use* (Wallingford, UK: CABI).

Maxted, N., Ford-Lloyd, B. V., Jury, S., Kell, S., and Scholten, M. (2006). Towards a definition of a crop wild relative. *Biodiversity Conserv.* 15 (8), 2673–2685. doi: 10.1007/s10531-005-5409-6

Maxted, N., Scholten, M., Codd, R., and Ford-Lloyd, B. (2007). Creation and use of a national inventory of crop wild relatives. *Biol. Conserv.* 140 (1-2), 142–159. doi: 10.1016/j.biocon.2007.08.006

Maxted, N., and Vincent, H. (2021). Review of congruence between global crop wild relative hotspots and centres of crop origin/diversity. *Genet. Resour. Crop Evol.* 68 (4), 1283–1297. doi: 10.1007/s10722-021-01114-7

Mertens, A., Swennen, R., Rønsted, N., Vandelook, F., Panis, B., Sachter-Smith, G., et al. (2021). Conservation status assessment of banana crop wild relatives using species distribution modelling. *Diversity Distributions* 27 (4), 729–746. doi: 10.1111/ddi.13233

Metwally, E., Sharshar, M., Masoud, A., Kilian, B., Sharma, S., Masry, A., et al. (2021). Development of high yielding cowpea [Vigna unguiculata (L.) walp.] lines with improved quality seeds through mutation and pedigree selection methods. *Horticulturae* 7 (9), 271. doi: 10.3390/horticulturae7090271

Meyer, A., and Barton, N. (2019). Botanic Gardens Are Important Contributors to Crop Wild Relative Preservation. *Crop Sci.* 59, 2404–12. doi: 10.2135/cropsci2019.06.0358

Mittler, R., and Blumwald, E. (2010). Genetic engineering for modern agriculture: challenges and perspectives. *Annu. Rev. Plant Biol.* 61, 443–462. doi: 10.1146/annurev-arplant-042809-112116

Moore, J. D., Kell, S. P., Iriondo, J. M., Ford-Lloyd, B. V., and Maxted, N. (2008). CWRML: representing crop wild relative conservation and use data in XML. *BMC Bioinf.* 9 (1), 1–7. doi: 10.1186/1471-2105-9-116

Mounce, R., Smith, P., and Brockington, S. (2017). Ex situ conservation of plant diversity in the world's botanic gardens. *Nat. Plants* 3 (10), 795–802. doi: 10.1038/s41477-017-0019-3

Nagatoshi, Y., and Fujita, Y. (2019). Accelerating soybean breeding in a CO2-supplemented growth chamber. *Plant Cell Physiol.* 60 (1), 77–84. doi: 10.1093/pcp/pcy189

Nduche, M., Brehm, J. M., Abberton, M., Omosun, G., and Maxted, N. (2021). "West African Crop wild relative checklist, prioritization and inventory," in *Genetic resources* 2(4), 55–65. doi: 10.46265/genresj.EIFL1323

Nelson, R., Wiesner-Hanks, T., Wisser, R., and Balint-Kurti, P. (2018). Navigating complexity to breed disease-resistant crops. *Nat. Rev. Genet.* 19 (1), 21–33. doi: 10.1038/nrg.2017.82

Ng'uni, D., Munkombwe, G., Mwila, G., Gaisberger, H., Brehm, J. M., Maxted, N., et al. (2019). Spatial analyses of occurrence data of crop wild relatives (CWR) taxa as tools for selection of sites for conservation of priority CWR in Zambia. *Plant Genet. Resour.* 17 (2), 103–114. doi: 10.1017/S1479262118000497

Obsie, E. Y., Qu, H., and Drummond, F. (2020). Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. *Comput. Electron. Agric.* 178, 105778. doi: 10.1016/j.compag.2020.105778

Papa, R., and Gepts, P. (2003). Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (Phaseolus vulgaris l.) from mesoamerica. *Theor. Appl. Genet.* 106 (2), 239–250. doi: 10.1007/s00122-002-1085-z

Parker, T. A., Berny Mier y Teran, J. C., Palkovic, A., Jernstedt, J., and Gepts, P. (2020). Pod indehiscence is a domestication and aridity resilience trait in common bean. *New Phytol.* 225 (1), 558–570. doi: 10.1111/nph.16164

Periyannan, S., Moore, J., Ayliffe, M., Bansal, U., Wang, X., Huang, L., et al. (2013). The gene Sr33, an ortholog of barley mla genes, encodes resistance to wheat stem rust race Ug99. *Science* 341 (6147), 786–788. doi: 10.1126/science.1239028

Perrino, E. V., and Perrino, P. (2020). Crop wild relatives: know how past and present to improve future research, conservation and utilization strategies, especially in Italy: a review. *Genet. Resour. Crop Evol.* 67 (5), 1067–1105. doi: 10.1007/s10722-020-00930-7

Perrino, E. V., and Wagensommer, R. P. (2022). Crop wild relatives (CWRs) threatened and endemic to Italy: Urgent actions for protection and use. *Biol. (Basel)* 11 (2), 193. doi: 10.3390/biology11020193

Phillips, J., Asdal, Å., Magos Brehm, J., Rasmussen, M., and Maxted, N. (2016). *In situ* and ex situ diversity analysis of priority crop wild relatives in Norway. *Diversity Distributions* 22 (11), 1112–1126. doi: 10.1111/ddi.12470

Phillips, J., Kyratzis, A., Christoudoulou, C., Kell, S., and Maxted, N. (2014). Development of a national crop wild relative conservation strategy for Cyprus. *Genet. Resour. Crop Evol.* 61 (4), 817–827. doi: 10.1007/s10722-013-0076-z

Pimentel, D., Wilson, C., McCullum, C., Huang, R., Dwen, P., Flack, J., et al. (1997). Economic and environmental benefits of biodiversity. *BioScience* 47 (11), 747–757. doi: 10.2307/1313097

Pironon, S., Borrell, J. S., Ondo, I., Douglas, R., Phillips, C., Khoury, C. K., et al. (2020). Toward unifying global hotspots of wild and domesticated biodiversity. *Plants* 9 (9):1128. doi: 10.3390/plants9091128

PolicyReport (2016) *In situ and ex situ conservation, two sides of the same coin*. Available at: https://www.cwrdiversity.org/wp/wp-content/uploads/2016/11/In-Situ-Ex-Situ-Policy-Brief.pdf.

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.

Postman, J., Hummer, K., Ayala-Silva, T., Bretting, P., Franko, T., Kinard, G., et al. (2010). GRIN-Global: An international project to develop a global plant genebank information management system. *Acta Hortic.* 859, 49–55. doi: 10.17660/ActaHortic.2010.859.4

Rahman, W., Brehm, J. M., Maxted, N., Phillips, J., Contreras-Toledo, A. R., Faraji, M., et al. (2021). Gap analyses of priority wild relatives of food crop in current ex situ and *in situ* conservation in Indonesia. *Biodiversity Conserv.* 30 (10), 2827–2855. doi: 10.1007/s10531-021-02225-4

Rana, M. M., Takamatsu, T., Baslam, M., Kaneko, K., Itoh, K., Harada, N., et al. (2019). Salt tolerance improvement in rice through efficient SNP marker-assisted selection coupled with speed-breeding. *Int. J. Mol. Sci.* 20 (10), 2585–2585. doi: 10.3390/ijms20102585

Ranganathan, J., Vennard, D., Waite, R., Dumas, P., Lipinski, B., and Searchinger, T. (2016). Shifting diets for a sustainable food future. *World Resour. Institute: Washington DC U.S.A.*

Ran, Y., Liang, Z., and Gao, C. (2017). Current and future editing reagent delivery systems for plant genome editing. *Sci. China Life Sci.* 60 (5), 490–505. doi: 10.1007/s11427-017-9022-1

Ratnayake, S. S., Kariyawasam, C. S., Kumar, L., Hunter, D., and Liyanage, A. S. U. (2021). Potential distribution of crop wild relatives under climate change in Sri Lanka: implications for conservation of agricultural biodiversity. *Curr. Res. Environ. Sustainability* 3, 100092. doi: 10.1016/j.crsust.2021.100092

Raubach, S., Kilian, B., Dreher, K., Amri, A., Bassi, F. M., Boukar, O., et al. (2021). From bits to bites: Advancement of the germinate platform to support prebreeding informatics for crop wild relatives. *Crop Sci.* 61 (3), 1538–1566. doi: 10.1002/csc2.20248

Raza, A., Razzaq, A., Mehmood, S. S., Zou, X., Zhang, X., Lv, Y., et al. (2019). Impact of climate change on crops adaptation and strategies to tackle its outcome: A review. *Plants* 8 (2), 34. doi: 10.3390/plants8020034

Rizal, G., Karki, S., Alcasid, M., Montecillo, F., Acebron, K., Larazo, N., et al. (2014). Shortening the breeding cycle of sorghum, a model crop for research. *Crop Sci.* 54, 520–529. doi: 10.2135/cropsci2013.07.0471

Robinson, A., Bell, A., Dinghe, N., and Stelly, D. (2004). "Status report on introgression of reniform nematode resistance from gossypium longicalyx," In: *Proceedings of the Beltwide Cotton Conferences*, San Antonio, Texas.

Rosenow, D. T., Dahlberg, J. A., Stephens, J. C., Miller, F. R., Barnes, D. K., Peterson, G. C., et al. (1997). Registration of 63 converted sorghum germplasm lines from the sorghum conversion program. *Crop Sci.* 37 (4), 1399–1400. doi: 10.2135/cropsci1997.0011183X003700040090x

Saintenac, C., Zhang, W., Salcedo, A., Rouse Matthew, N., Trick Harold, N., Akhunov, E., et al. (2013). Identification of wheat gene Sr35 that confers resistance to Ug99 stem rust race group. *Science* 341 (6147), 783–786. doi: 10.1126/science.1239022

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun.* 11 (1), 4572. doi: 10.1038/s41467-020-18404-w

Saxena, R. K., Edwards, D., and Varshney, R. K. (2014). Structural variations in plant genomes. *Briefings Funct. Genomics* 13 (4), 296–307. doi: 10.1093/bfgp/elu016

Schouten, H. J., Tikunov, Y., Verkerke, W., Finkers, R., Bovy, A., Bai, Y., et al. (2019). Breeding has increased the diversity of cultivated tomato in the Netherlands. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.01606

Schwacke, R., Ponce-Soto, G. Y., Krause, K., Bolger, A. M., Arsova, B., Hallab, A., et al. (2019). MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol. Plant* 12 (6), 879–892. doi: 10.1016/j.molp.2019.01.003

Shan-e-Ali Zaidi, S., and Mansoor, S. (2017). Viral vectors for plant genome engineering. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00539/bibtex

Shaw, P. D., Raubach, S., Hearne, S. J., Dreher, K., Bryan, G., McKenzie, G., et al. (2017). Germinate 3: development of a common platform to support the distribution of experimental data on crop wild relatives. *Crop Sci.* 57 (3), 1259–1273. doi: 10.2135/cropsci2016.09.0814

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of brassica napus. *Nat. Plants* 6 (1), 34–45. doi: 10.1038/s41477-019-0577-7

Stephens, J. C., Miller, F. R., and Rosenow, D. T. (1967). Conversion of alien sorghums to early combine Genotypes1. *Crop Sci.* 7 (4), 396–396. doi: 10.2135/cropsci1967.0011183X000700040036x

Steuernagel, B., Periyannan, S. K., Hernández-Pinzón, I., Witek, K., Rouse, M. N., Yu, G., et al. (2016). Rapid cloning of disease-resistance genes in plants using mutagenesis and sequence capture. *Nat. Biotechnol.* 34 (6), 652–655. doi: 10.1038/nbt.3543

Stoilova, T., van Zonneveld, M., Roothaert, R., and Schreinemachers, P. (2019). Connecting genebanks to farmers in East Africa through the distribution of

vegetable seed kits. *Plant Genet. Resources: Characterization Utilization* 17 (3), 306–309. doi: 10.1017/S1479262119000017

Summers, R. W., and Brown, J. K. M. (2013). Constraints on breeding for disease resistance in commercially competitive wheat cultivars. *Plant Pathol.* 62 (S1), 115–121. doi: 10.1111/ppa.12165

Tao, Y., Luo, H., Xu, J., Cruickshank, A., Zhao, X., Teng, F., et al. (2021). Extensive variation within the pan-genome of cultivated and wild sorghum. *Nat. Plants* 7 (6), 766–773. doi: 10.1038/s41477-021-00925-x

Tas, N., West, G., Kircalioglu, G., Topaloglu, S. B., Phillips, J., Kell, S., et al. (2019). Conservation gap analysis of crop wild relatives in Turkey. *Plant Genet. Resour.* 17 (2), 164–173. doi: 10.1017/S1479262118000564

Tay Fernandez, C. G., Nestor, B. J., Danilevicz, M. F., Gill, M., Petereit, J., Bayer, P. E., et al. (2022). Pangenomes as a resource to accelerate breeding of under-utilised crop species. *Int. J. Mol. Sci.* 23 (5), 2671. doi: 10.3390/ijms23052671

Taylor, N., Holubec, V., Chobot, K., Parra-Quijano, M., Maxted, N., and Kell, S. (2013). Systematic crop wild relative conservation planning for the Czech republic. *Crop Wild relative* 9, 5–9. doi: 10.1079/9781845930998.000

Teso, M. L. R., Lamas, E. T., Parra-Quijano, M., de la Rosa, L., Fajardo, J., and Iriondo, J. M. (2018). National inventory and prioritization of crop wild relatives in Spain. *Genet. Resour. Crop Evol.* 65 (4), 1237–1253. doi: 10.1007/s10722-018-0610-0

Tyack, N., Dempewolf, H., and Khoury, C. K. (2020). The potential of payment for ecosystem services for crop wild relative conservation. *Plants* 9 (10), 1305. doi: 10.3390/plants9101305

Van Bel, M., Silvestri, F., Weitz, E. M., Kreft, L., Botzki, A., Coppens, F., et al. (2022). PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. *Nucleic Acids Res.* 50 (D1), D1468–D1474. doi: 10.1093/nar/gkab1024

van Bemmelen van der Plaat, A., van Treuren, R., and van Hintum, T. J. L. (2021). Reliable genomic strategies for species classification of plant genetic resources. *BMC Bioinf.* 22 (1), 173. doi: 10.1186/s12859-021-04018-6

Vavilov, N. (1926). Center of origin of cultivated plants. *Papers Appl. Botany Genet. Plant Breeding* 16, 1–248.

Vavilov, N. I., Vavylov, M. I., and Dorofeev, V. F. (1992). *Origin and geography of cultivated plants* (Cambridge: Cambridge University Press).

Vendelbo, N. M., Mahmood, K., Steuernagel, B., Wulff, B. B. H., Sarup, P., Hovmøller, M. S., et al. (2022). Discovery of resistance genes in rye by targeted long-read sequencing and association genetics. *Cells* 11 (8). doi: 10.3390/cells11081273

Vincent, H., Wiersema, J., Kell, S., Fielder, H., Dobbie, S., Castañeda-Álvarez, N. P., et al. (2013). A prioritized crop wild relative inventory to help underpin global food security. *Biol. Conserv.* 167, 265–275. doi: 10.1016/j.biocon.2013.08.011

Vincent, H., Hole, D., and Maxted, N. (2022). Congruence between global crop wild relative hotspots and biodiversity hotspots. *Biological Conservation* 265, 109432. doi: 10.1016/j.biocon.2021.109432

Viruel, J., Kantar, M. B., Gargiulo, R., Hesketh-Prichard, P., Leong, N., Cockel, C., et al. (2021). Crop wild phylorelatives (CWPs): phylogenetic distance, cytogenetic compatibility and breeding system data enable estimation of crop wild relative gene pool classification. *Botanical J. Linn. Soc.* 195 (1), 1–33. doi: 10.1093/botlinnean/boaa064

Wambugu, P. W., Ndjiondjop, M.-N., and Henry, R. J. (2018). Role of genomics in promoting the utilization of plant genetic resources in genebanks. *Briefings Funct. Genomics* 17 (3), 198–206. doi: 10.1093/bfgp/ely014

Wang, M., Yang, J., Wan, J., Tao, D., Zhou, J., Yu, D., et al. (2020). A hybrid sterile locus leads to the linkage drag of interspecific hybrid progenies. *Plant Divers.* 42 (5), 370–375. doi: 10.1016/j.pld.2020.07.003

Watson, A., Ghosh, S., Williams, M. J., Cuddy, W. S., Simmonds, J., Rey, M.-D., et al. (2018). Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* 4, 23–29. doi: 10.1038/s41477-017-0083-8

Wegrzyn, J. L., Liechty, J. D., Stevens, K. A., Wu, L.-S., Loopstra, C. A., Vasquez-Gross, H. A., et al. (2014). Unique features of the loblolly pine (Pinus taeda l.) megagenome revealed through sequence annotation. *Genetics* 196 (3), 891–909. doi: 10.1534/genetics.113.159996

Wilkinson, P. A., Allen, A. M., Tyrrell, S., Wingen, L. U., Bian, X., Winfield, M. O., et al. (2020). CerealsDB–new tools for the analysis of the wheat genome: update 2020. *Database* 2020, 1–13. doi: 10.1093/database/baaa060

Witek, K., Jupe, F., Witek, A. I., Baker, D., Clark, M. D., and Jones, J. D. G. (2016). Accelerated cloning of a potato late blight-resistance gene using RenSeq and SMRT sequencing. *Nat. Biotechnol.* 34 (6), 656–660. doi: 10.1038/nbt.3540

Xiang, Z., Chen, Y., Chen, Y., Zhang, L., Liu, M., Mao, D., et al. (2022). Agrobacterium-mediated high-efficiency genetic transformation and genome editing of chaling common wild rice (Oryza rufipogon griff.) using scutellum tissue of embryos in mature seeds. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.849666

Xu, X., Liu, X., Ge, S., Jensen, J. D., Hu, F., Li, X., et al. (2012). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30 (1), 105–111. doi: 10.1038/nbt.2050

Yahiaoui, N., Kaur, N., and Keller, B. (2009). Independent evolution of functional Pm3 resistance genes in wild tetraploid wheat and domesticated bread wheat. *Plant J.* 57 (5), 846–856. doi: 10.1111/j.1365-313X.2008.03731.x

Yik, C. P., and Birchfield, W. (1984). Resistant germplasm in gossypium species and related plants to rotylenchulus reniformis. *J. Nematol* 16 (2), 146–153.

Yocca, A. E., and Edger, P. P. (2022). Machine learning approaches to identify core and dispensable genes in pangenomes. *Plant Genome* 15 (1), e20135. doi: 10.1002/tpg2.20135

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., et al. (2021). A route to *de novo* domestication of wild allotetraploid rice. *Cell* 184 (5), 1156–1170. doi: 10.1016/j.cell.2021.01.013

Zair, W., Maxted, N., and Amri, A. (2018). Setting conservation priorities for crop wild relatives in the fertile crescent. *Genet. Resour. Crop Evol.* 65 (3), 855–863. doi: 10.1007/s10722-017-0576-3

Zair, W., Maxted, N., Brehm, J. M., and Amri, A. (2021). Ex situ and *in situ* conservation gap analysis of crop wild relative diversity in the fertile crescent of the middle East. *Genet. Resour. Crop Evol.* 68 (2), 693–709. doi: 10.1007/s10722-020-01017-z

Zhang, H., Demirer, G. S., Zhang, H., Ye, T., Goh, N. S., Aditham, A. J., et al. (2019). DNA Nanostructures coordinate gene silencing in mature plants. *Proc. Natl. Acad. Sci.* 116 (15), 7543–7548. doi: 10.1073/pnas.1818290116

Zhan, X., Lu, Y., Zhu, J. K., and Botella, J. R. (2021). Genome editing for plant research and crop improvement. *J. Integr. Plant Biol.* 63 (1), 3–33. doi: 10.1111/jipb.13063

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33 (4), 408–414. doi: 10.1038/nbt.3096

Zhu, H., Li, C., and Gao, C. (2020). Applications of CRISPR-cas in agriculture and plant biotechnology. *Nat. Rev. Mol. Cell Biol.* 21, 661–677. doi: 10.1038/s41580-020-00288-9

Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., et al. (2018). *De novo* domestication of wild tomato using genome editing. *Nat. Biotechnol.* 36 (12), 1211–1216. doi: 10.1038/nbt.4272

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 08 December 2022
DOI 10.3389/fpls.2022.1008666

OPEN ACCESS

EDITED BY
Jinyoung Y. Barnaby,
United States Department of Agriculture (USDA), United States

REVIEWED BY
Jorge Carlos Berny Mier y Teran,
University of California, Davis, United States
Jagadish Rane,
Indian Council of Agricultural Research (ICAR), India
Barbara Pipan,
Agricultural institute of Slovenia, Slovenia

*CORRESPONDENCE
Milan Oldřich Urban
m.urban@cgiar.org
Diego Felipe Conejo Rodriguez
d.conejo@cgiar.org

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

RECEIVED 01 August 2022
ACCEPTED 14 November 2022
PUBLISHED 08 December 2022

CITATION
Rodriguez DFC, Urban MO, Santaella M, Gereda JM, Contreras AD and Wenzl P (2022) Using phenomics to identify and integrate traits of interest for better-performing common beans: A validation study on an interspecific hybrid and its Acutifolii parents.
Front. Plant Sci. 13:1008666.
doi: 10.3389/fpls.2022.1008666

Using phenomics to identify and integrate traits of interest for better-performing common beans: A validation study on an interspecific hybrid and its Acutifolii parents

Diego Felipe Conejo Rodriguez [1*], Milan Oldřich Urban [2*], Marcela Santaella [1], Javier Mauricio Gereda [1], Aquiles Darghan Contreras [3] and Peter Wenzl [1]

[1]Genetic Resources Program, International Center for Tropical Agriculture (CIAT), Recta Cali-Palmira, Valle del Cauca, Colombia, [2]Bean Physiology and Breeding Program, International Center for Tropical Agriculture, Recta Cali-Palmira, Valle del Cauca, Colombia, [3]Department of Agronomy, Faculty of Agricultural Sciences, Universidad Nacional de Colombia, Bogotá, Colombia

Introduction: Evaluations of interspecific hybrids are limited, as classical genebank accession descriptors are semi-subjective, have qualitative traits and show complications when evaluating intermediate accessions. However, descriptors can be quantified using recognized phenomic traits. This digitalization can identify phenomic traits which correspond to the percentage of parental descriptors remaining expressed/visible/measurable in the particular interspecific hybrid. In this study, a line of *P. vulgaris*, *P. acutifolius* and *P. parvifolius* accessions and their crosses were sown in the mesh house according to CIAT seed regeneration procedures.

Methodology: Three accessions and one derived breeding line originating from their interspecific crosses were characterized and classified by selected phenomic descriptors using multivariate and machine learning techniques. The phenomic proportions of the interspecific hybrid (line INB 47) with respect to its three parent accessions were determined using a random forest and a respective confusion matrix.

Results: The seed and pod morphometric traits, physiological behavior and yield performance were evaluated. In the classification of the accession, the phenomic descriptors with highest prediction force were Fm', Fo', Fs', LTD, Chl, seed area, seed height, seed Major, seed MinFeret, seed Minor, pod AR, pod Feret, pod round, pod solidity, pod area, pod major, pod seed weight and pod weight. Physiological traits measured in the interspecific hybrid present 2.2% similarity with the *P. acutifolius* and 1% with the *P. parvifolius* accessions. In addition, in seed morphometric characteristics, the hybrid showed 4.5% similarity with the *P. acutifolius* accession.

Conclusions: Here we were able to determine the phenomic proportions of individual parents in their interspecific hybrid accession. After some careful generalization the methodology can be used to: i) verify trait-of-interest transfer from *P. acutifolius* and *P. parvifolius* accessions into their hybrids; ii) confirm selected traits as "phenomic markers" which would allow conserving desired physiological traits of exotic parental accessions, without losing key seed characteristics from elite common bean accessions; and iii) propose a quantitative tool that helps genebank curators and breeders to make better-informed decisions based on quantitative analysis.

KEYWORDS

phenomic descriptors, phenomic proportions, interspecific hybrid, image analysis, machine learning

Introduction

Genebank plant genetic resources comprise the representative diversity of genetic material contained in traditional varieties and modern cultivars, as well as in the crop wild relatives and other wild plant species that can be used now or in the future for food and agriculture (Wang and Zhang, 2011). Currently, there are about 1,750 genebanks worldwide that conserve 7.4 million accessions of agricultural genetic resources (Noriega et al., 2019). Eleven CGIAR genebanks conserve about 730,000 accessions among crops, trees, and forages, of which the International Center for Tropical Agriculture (CIAT) conserves 37,987 bean accessions, 23,140 forage accessions and nearly 6,000 Cassava accessions (Noriega et al., 2019). Despite this great diversity, only approximately 10% of the accessions from the 1,750 genebanks is used in plant breeding, mainly because of poor phenotypic and genotypic characterization or lack of agronomic traits evaluation (de Carvalho et al., 2013; Tadesse et al., 2019; Nguyen and Norton, 2020; Kholova et al., 2022).

P. acutifolius (tepary bean) is an important species in common bean breeding, due to its adaptation to abiotic and biotic stress (Singh and Munoz, 1999; Porch et al., 2013; Kusolwa et al., 2016). The use of cultivated and wild relatives of *P. vulgaris* by the common bean breeding program at CIAT started in the 1980s, with the aim of generating lines with elevated levels of introgression from *P. acutifolius* and/or *P. parvifolius* ~ *P. montanus* (Debouck, 2021), using techniques such as congruity backcrossing (CBC) and recurrent backcrossing (RBC) (Haghighi and Ascher, 1988; Mejía-Jiménez et al., 1994; Singh et al., 1998) with the help of bridge genotypes.

Classification of hybrids based on phenotypic traits was done in the 1960s (Allendorf et al., 2001), however, the detection of morphological traits usually assumes that hybrids are phenotypically intermediate to the parents. This is often not the case, because hybrids express a mosaic of parental phenotypes (Arnold, 1997) influenced also by environmental conditions. Furthermore, morphological characters do not allow determining whether an individual is a first-generation hybrid (F1), a backcross or late generation hybrid (Allendorf et al., 2001).

Recently, there have been published studies that promote characterization processes using phenomic descriptors. When compared with conventional descriptors, these showed a better capacity for analysis of phenotypic variability (Rosero et al., 2019; Nankar et al., 2020). Phenomic descriptors has both qualitative and quantitative characters and deal with agronomic, morphological, physiological, and colorimetric traits of accessions which are captured by proximal sensors such as cameras, fluorometers, trichromatic, multispectral and hyperspectral sensors. Phenomic descriptors have a "high-throughput" character of data, which means, hundreds of accessions can possibly be characterized/screened in a reasonable time. However, the sheer volume, variety and veracity of imagery and remote-sensing data still present limits in data analysis (Singh et al., 2016).

To solve this problem, there are currently several machine learning models, such as partial least squares (PLS), random forest (RF), support vector machines (SVM), and neural networks (NN) used in phenomics data analytics (Araus et al., 2022). Based on phenomic traits, classification of accessions and their hybrids has been attempted. This has been made possible with the development of phenomic and machine learning methods with thousands of data points from each individual, (Parmley et al., 2019; Soltis et al., 2020; Feldmann et al., 2020; Henao-Rojas et al., 2021).

In this work, we propose a Phaseolus-oriented methodology for the detection of phenomic proportions of interspecific hybrids with respect to their parents. We used multivariate and machine learning methods to characterize and classify

three parental line accessions (a cultivated *P. vulgaris*, a domesticated *P. acutifolius*, and a wild *P. parvifolius* - P. *montanus*) with its interspecific hybrid accession.

The phenomic proportions correspond to the percentage of parental descriptors which remain expressed/visible/measurable in the particular interspecific hybrid. Correspondingly, in this study, phenomic proportions show the phenomic traits portions quantified in the three parental lines and verified in an interspecific hybrid. We hope this methodology will provide a first step to help genebank curators, breeders, physiologists and others to i) make detailed quantitative comparisons of selected phenomic traits between accessions of interest, and ii) better manage and understand their genetic resources. After some generalization using other parents/hybrid collections or random selection, this methodology could facilitate a deeper understanding about i) crossings, heritability and breeding success; ii) functional trait diversity; iii) species domestication/evolution and genetic recombination; and iv) how to substantially increase genetic gains (in tandem with genomics).

Methodology

Plant material and experimental design

Materials used in this study were *P. acutifolius* (G40001 – CIAT genebank accession number), *P. parvifolius - P. montanus* (G40102), *P. vulgaris* (G5773) and their interspecific cross hybrid line INB 47 (G52443), all obtained from the bean collection at CIAT Genebank. Accessions G40001 (*P. acutifolius*) and G40102 (*P. parvifolius - P. montanus*) display a type III growth habit (indeterminate prostrate growth); while accessions G5773 and hybrid G52443 exhibit a type I growth habit (determinate bush growth; growth habits defined according to Fernández de Córdoba et al., 1991). Accession G40001 showed heat and drought resistance (Mejía-Jiménez et al., 1994) and G40102 was highly resistant to common bacterial blight (CBB) (Singh et al., 1998).

The interspecific hybridization was a product of artificial crossings, carried out by the CIAT common bean breeding program, from crosses between a popular commercial bean variety Ica Pijao, *P. parvifolius* (G40102) and an interspecific line, with five cycles of congruity backcrossing (CBC_5) between Ica Pijao and *P. acutifolius* (G40001) (Mejía-Jiménez et al., 1994). The pedigree of the interspecific hybrid line is as follows: INB 47 (G52443) = ICA PIJAO x (G40102 x (ICA PIJAO x (G40102 x (ICA PIJAO x (ICA PIJAO x (ICA PIJAO x G40001; CBC_5).

The INB 47 line was developed from ten (10) cycles of selection pressure based on commercial characteristics including growth habit, seed type and yield in the experimental station in Santander de Quilichao and in Palmira (both sites in Colombia), CIAT (Personal communication, Common bean breeding program, CIAT). The selection was mainly focused on conserving the commercial seed type similar to *P. vulgaris* – Ica Pijao, while introducing resistance to bacterial diseases (Mejía-Jiménez et al., 1994). The phenotypic characteristics of the parental lines and the interspecific hybrid are shown in Figure 1. These characteristics were observed during the experiment at the regeneration station of CIAT's genetic resources program (GRP) in Palmira.

Our experiments were performed in two periods: from October 2018 to January 2019, and from January 2019 to April 2019, in a mesh house at CIAT, Palmira, Colombia (3°30′17″ N, 76°21′24″ W, 950 masl). The cultivation protocol was used according to standard operating procedures used in CIAT genebanks when multiplying accessions. The experimental conditions inside the mesh house presented: i) an average daily temperature of 38°C, with a daily maximum temperature of 41°C at midday hours, and daily minimal temperature of 27°C; ii) a minimum relative humidity of 31% and a maximum of 65%; and with iii) an average photosynthetically active radiation (PAR) of 1680 µmol m-2 s-1 during the sampling of physiological variables (please, explore variables at: https://photosynq.org/projects/domestication-syndrome/explore); iv) 12 h of natural light, and v) conventional agronomic management and fertigation using drip irrigation. The plants were sown in a substrate of coconut fiber substrate for hydroponic systems (120 × 40 × 40 cm), composed of 100% peat, which is highly efficient in conserving water and guarantees health development in the early stages of cultivation (elaborated in Spain by the company Berger, https://www.berger.ca/en/horticultural-products/). Water irrigation and fertilization scenarios were the same for all plants (Supplementary Material Table 1 - Fertilization).

Each plant was tutored with agricultural mesh, starting at germination, to minimize human intervention in the span of plant development and guarantee adequate growth. The experimental design was a complete randomized block design, in which each accession acted as a treatment and each plant as an experimental unit. Five (5) plants per accession were planted, and three (3) independent technical repetitions measured in each period. To evaluate differences between accessions with special focus on the hybrid, we decided to compare morphometric, physiological, and agronomic traits.

Morphometric descriptors

We evaluated the morphometric aspects of pod shape and lateral seed shape. Seed morphological characterization was carried out using phenomic descriptors based on image analysis, including the following traits: seed Area (cm^2), Perimeter (cm), Width (cm), Height (cm), Major, Minor, MinorFeret, MajorFeret, Aspect Radio (AR), Circularity, Roundness, and Solidity according to Schlautman et al. (2020)

and Rosero et al. (2019) (see details in Supplementary Material Table 2 – Morphometric descriptions).

The digital images of each of the accession's pod/seed were obtained using Canon SX60 HS camera with a digital resolution of 16.1 megapixels, an image area of 1080 × 1080 pixels in an automatic format. We captured images in a RAW format to ensure maximum image quality. Images were taken separately for seeds and pods. Each picture was processed using the DCRAW plugin of ImageJ 5.4 (https://imagej.net/software/fiji/).

After plant harvest, images were captured from pods and seeds of every accession, considering the level of luminosity and the contrast of the background. To capture color of the seeds, we used a contrasting background and standardized color scale 24ColorCard Camera Trax Card-3x5 (Supplementary Material Figure 1 – Color scale and photobox used). Images of the pod shape and lateral shape of the seeds were processed using ImageJ software (Eliceiri et al., 2012), following the protocol: (I) Settlement of the scale (pixels to cm), (II) Image binarization using the Max Entropy and Huang methods (Huang and Wang, 1995), (III) Definition of regions of interest (ROI) from pod and seed selections, and (IV) Extraction of morphometric measurements of each of the selected ROIs (Supplementary Material Figure 2 – imagen analysis process).

Physiological descriptors

For physiological measurements, we used the MultispeQ device (Kuhlgert et al., 2016; PhotosynQ, USA). The device includes climatic and plant variables that facilitate characterizing the physiological performance of plants in their environment. All data were captured at midday between 11:00 am and 1:00 pm, with the aim of comparing the data collected under more stable temperature and lower air humidity (RH) under mesh house conditions. Measurements were taken three times a week in three technical (plant as experimental unit) repetitions per accession in all three replicates and both periods. In total, 1,022 MultispeQ observations were captured for the four accessions. Samplings were carried out every 15 days from December 22, 2018 to April 23, 2019. Samples were taken during phenological stages 22 to 85 according to the BBCH scale from plant branching until harvest maturity (Feller et al., 1995). The classical protocol was used: Leaf Photosynthesis MultispeQ V1.0 (the raw data are available at: https://photosynq.org/projects/domestication-syndrome; ID 5685).

Briefly, the MultispeQ proximal sensors measure photosynthetic parameters including: i) quantum yield of photosystem II (ΦII – Phi2); ii) non-photochemical quenching (ΦNPQ - PhiNPQ); iii) energy losses for heat dissipation (ΦNO - PhiNO); iv) relative chlorophyll (Chl); v) linear electron flux (LEF); vi) leaf temperature differential (LTD); vii) maximum variable fluorescence at a steady-state conditions (Fm'); viii) minimum variable fluorescence during dark phase after a steady-state (Fo'); ix) variable fluorescence at a steady-state conditions (Fs'); x) efficiency of open reactions centers in the light (Fv'/Fm'); xi) fraction of open PSII centers when QA is oxidized (qL); xii) photochemical quenching relating PSII maximum efficiency (qP); xiii) fluorescence decrease ratio (RFd), and xiv) leaf thickness (Kuhlgert et al., 2016; Fernández-Calleja et al., 2020; Deva et al., 2020). In addition, the device also measures environmental conditions like light intensity (photosynthetically active radiation, PAR), air temperature and air humidity (see Supplementary Material – Figure 3 for temperature range, humidity and PAR). The data acquired can be visualized on the PhotosynQ platform (i.e. exploratory analysis of the data). Thus, at air temperatures above 32°C, photosynthetic activity is restricted in common bean of determinate growth type (Beebe et al., 2011; Deva et al., 2020).

Yield components

To calculate seed yield, ten pods were taken randomly from each of three similar plants per accession during the final harvest at BBCH 89. Seeds were pre-dried according to the Genetic Resources Program methodology (Salazar et al., 2020). Seed weight was measured with high precision scales with seed average humidity of 14%. The dry weight of seed at harvest (PSW), weight of pod with seed at harvest (PW), dry weight of pod without seed (pod walls) (VW), number of seeds (SN) and the harvest index at the pod level (PHI = ((Dry weight of seeds at harvest)/(Dry weight of whole pod at harvest))x100) were determined.

Multivariable analysis and Phaseolus accessions classification

Initially, we performed an outlier detection test using the Dobin library of R software (Kandanaarachchi and Hyndman, 2021). Principal Component Analysis (PCA) was carried out first on the parents, and the Principal Components (PC) with the highest contribution to the explained variance were extracted in each characterization group for each variable (performed using the FactorExtra library of R; Kassambara and Mundt, 2020). Predictor analysis was performed by random forests (randomForest library of R; Liaw and Wirner, 2018). Classification was performed on 100 trees, using 70% of the data for tree training, and 30% of the data for validation. For evaluating the classification model prediction with "out of bag" accuracy (OOB accuracy). The OOB is an error estimation technique used to evaluate the accuracy of the random forest (Janitza and Hornung,

2018). The OOB estimates accuracy across all classes (values above 1 - 10% are estimated as high accuracy; Kennedy et al., 2015).

The evaluation metric and confusion matrix were determined to observe the phenomic proportions of parent classification for each characterization group. The descriptors selected for each characterization component are determined by mean decrease in accuracy and the gini decrease index as parameters for feature selection. Analyses were run in R using the library "randomForestExplainer" and "caret" library (Paluszynska, 2017).

Phenomic proportions of the interspecific hybrid with respect to its parents

Initially, the contributions of the PCs from the classification of the parent lines were used for weighting the phenomic descriptors of the interspecific hybrid. Subsequently, the weighted phenomic descriptor values of both parents and the interspecific hybrid are standardized to values between 0 - 1. The phenomic proportions are determined from classifying parents and an interspecific hybrid using random forests. Phenomic descriptors of importance in the classification will be determined for each characterization component using the gini index and mean decrease accuracy. The prediction of the confusion matrix in the interspecific hybrid will be considered as the phenomic proportions that it presents with respect to each of its parents. A confusion matrix is typically created representing the summary of the number of correct and incorrect prediction results broken down by each parental line.

Finally, a non-parametric multivariate analysis of variance (MANOVA) was performed to determine if there are significant differences between the parents and their hybrids in each of the characterization components with the already prioritized descriptors. In our study we used MANOVA developed by Friedrich and Pauly (2018) which allows flexibility of normality assumptions and incorporates general heteroscedastic designs and potentially singular covariance matrices. It also improves the performance of small samples through bootstrap techniques. The analysis was performed using 10,000 iterations, modified ANOVA-type statistics (MATS), and the p-resampling value was determined from the parametric bootstrap approach (paramBS). MANOVA was performed for each characterization component separately (physiology, pod morphometry, seed morphometry and yield). In order to observe the significant differences between parents and its hybrid, the *post hoc* Tukey multivariate test was performed. This was done using the MANOVA.RM library of the free software R. The summary of the data analysis procedure can be seen in Figures 2, 3.

Results

The three parental accessions used in this study show contrasting phenotypic differences in morphological characteristics of the leaf shape, growth habit, flower shape and color, and seed size (Figure 1). *P. acutifolius* (G40001) (Figure 1A) presents an oval-lanceolate leaf shape and acute angles (less than 90 degrees) in conditions of high light intensity at midday hours. The flowers and seeds of *P. acutifolius* are white, with a straight pod shape, and a predominantly round-oval seed shape. *P. parvifolius* (G40102) (Figure 1B) is a wild accession with a lobed leaf shape, a light-purple flower color, and dark-purple and curved pods. *P. parvifolius* seeds are small, black to mottled grayish black, with a flattened truncated shape. *P. vulgaris* - Ica Pijao (G52443) (Figure 1C) has an oval leaf shape, dark purple flowers, mottled purple pods with slightly curved pod shape, and black seeds with a flattened oval shape. The interspecific hybrid (G52443 - INB 47) (Figure 1D) has a lanceolate leaf shape and acute leaf angles under high light intensity. It has purple flowers, with some flowers showing malformations in the floral wings. The pods are slightly curved with mottled purple colors, the seeds are black, and the seed shape is round cuboid.

The classification of the parent accessions using random forests, determined as predictive descriptors for seed shape were: Major, Area, Minot, Height and MinFeret; while in pod morphometry the predictive descriptors were: Major, Feret, Round, Aspect Radio (AR), Solidity and Area, presenting corresponding values of gini index higher than 40 and accuracy decrease 0.06 (Supplementary Material Figures 4AS, 4CS). The physiological variables, the phenomic descriptors Fo', LTD, Fs' and Fm', and in the yield components VW (Valve weight), PW (Pod weight), SPW (Seed pod weight) and PHI (Pod harvest index), presented gini index values higher than 35 and accuracy decrease higher than 0.075 (Supplementary Material Figures 4AS, 4CS). The table of the contributions of the PCs to each of the predictive descriptors is shown in Supplementary Material Table 3. The classification of the parents can be seen in the confusion matrix (Supplementary Material Figure 5 Matrix confusion parents' accessions).

Phenomic proportions of the hybrid respect to its parents

The low OOB value of 5.68% differences in the phenomic shape descriptors of seed of hybrid with respect to its parents. Interestingly, in the case of pod morphometric descriptors and physiological descriptors, the OOB values reached 22.44% and 36.76%, respectively; while 8.49% for the yield component (Supplementary Material Table 4 - OOB error).

The phenomic descriptors of importance in the classification of the parental accessions and the hybrid are presented in Figure 4. Generally, the descriptors of seed morphometry, Area and Minor were the most important, while in pod morphometry the descriptors Major, Feret and AR presented accuracy higher than 0.250 (Figures 4A, C). In the physiological descriptors, Fm' and Chl are the most important, and in the yield components the descriptors PHI and PW are the ones that presented the highest values of accuracy decrease with values higher than 0.125 (Figures 4B, D).

Using the predictions in the confusion table, the phenomic proportions of each of the characterization components were determined (Figure 5). The relationships of the predictions of the hybrid with its parents are most closely related to the common bean parent *P. vulgaris* - Ica Pijao (G52443). The values of proportions are as follows: 5.2 % for physiological traits, 9.8% for pod morphometric traits, and 4.1% for yield components.

Despite the hybrid's high relatedness to its *P. vulgaris* parent accession (the effect of multiple back-crosses), the hybrid presents phenomic proportions of 2.2% also with *P. acutifolius* (G40001) and 1% with *P. parvifolius* (G40102) accessions in physiological descriptors (Figure 5B), while 4.5% with *P. acutifolius* in seed morphometrics (Figure 5A), 0.6% for yield and 4.9% for pod morphometrics indicating successfully inherited traits from these parents as well (Figure 5C).

Phenomic proportions of the physiological components showed trait discrimination (difference) between the parents in traits like the Chl, Fm', Fo', LTD and Fs', respectively. Furthermore (Figure 4B), it was observed that the parental *P. parvifolius* shows contrasting physiological behavior when compared with the other two parents. This can be explained by its wild origin and different morphometric characteristics. The interspecific hybrid is closely related to the *P. vulgaris* accession and also the *P. acutifolius* accession, indicating that it conserves physiological characteristics mainly from these two parents. However, the hybrid presents higher phenomic proportions of the *P. vulgaris* accession and not of the *P. acutifolius* accession (Figure 5B). For seed morphometry, the

FIGURE 1
Phenotypic characteristics of the *Phaseolus* lines and interspecific hybrid, from its initial stages until harvest. **(A)** *P. acutifolius* (G40001), **(B)** *P. parvifolius* (G40102), **(C)** *P. vulgaris* - ICA Pijao (G5773) and **(D)** Interspecific hybrid (G52443 - INB 47). The figure shows the variation of morphological traits of the common bean accessions evaluated.

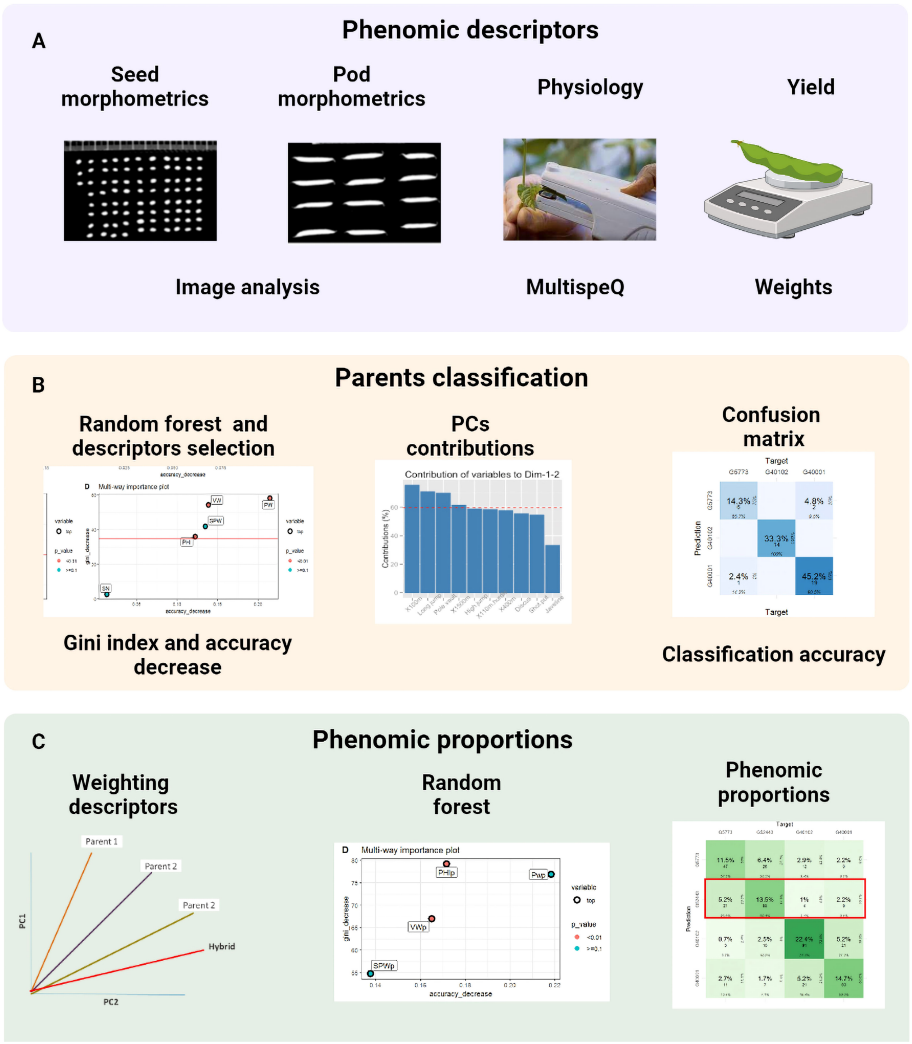


FIGURE 2
The procedure consists of three steps: **(A)** capture and processing of phenomic descriptors, **(B)** classification of parents using random forest and PCA and **(C)** phenomic ratios based on the weighting of the descriptors with the contribution of the PCs and phenomic ratios as the prediction of the hybrid with respect to its parents in the confusion table. This figure shows the different stages during the characterization process from data capture to data analysis.

confusion matrix clearly separates the interspecific hybrid and the parents with high precision, showing the hybrid has unique characteristics in seed morphometry, despite sharing 4.5% of phenomic proportion with *P. acutifolius* (Figure 5A). In pod morphometry, it is clear that there is a difference of our hybrid regarding its *P. acutifolius* and *P. parvifolius* parental accessions (Figure 5C). This supports the fact that there is no clear trait separation between the *P. vulgaris* parental and the interspecific hybrid and again is likely an influence of multiple back-crossing and/or environment-based limitations of *P. acutifolius*-related traits.

The data after MANOVA fitting, supported the rejection of the statistical hypothesis associated with shared characteristics between the hybrid and its parents (Table 1). In the physiological characterization (MultispeQ data), there were no significant differences with the parentals *P. acutifolius* and *P. vulgaris* accession; while with *P. parvifolius* accession differences were highly significant (< 0.001) (Table 1). In pod morphometry, the interspecific hybrid showed no differences with *P. acutifolius* and *P. vulgaris* accessions. In seed morphometry, the hybrid showed no differences with the parental *P. vulgaris*, while in the yield components it showed no differences with *P. parvifolius* and *P. vulgaris* (Table 1) accessions. The MANOVA supports the statistical differences of the hybrid and its parents in each characterization component, contrasting with those obtained in the random forest analysis.

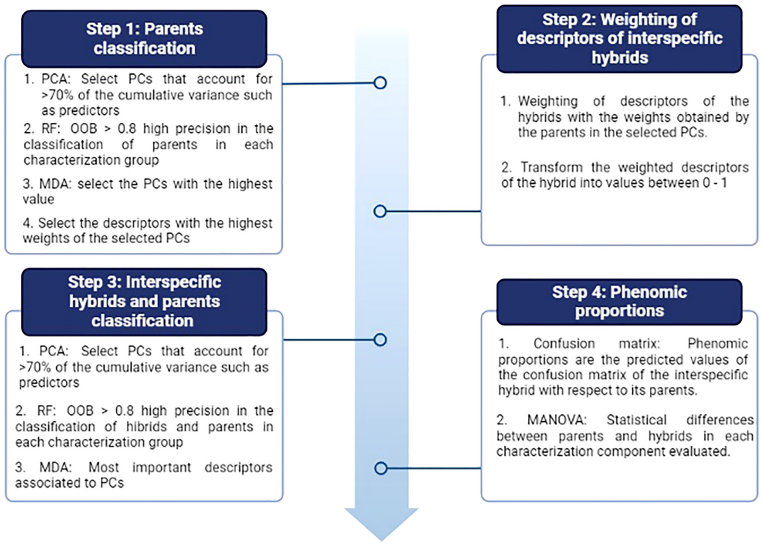


FIGURE 3
Procedure for the analysis of the phenomic proportions between interspecific hybrids and their parents. Four stages are observed that contemplate multivariate (PCA) and random forest (RF) analyses. OOB: Out of bag accuracy and MDA: Mean decrease accuracy. The figure shows the evaluation metrics used during each stage.

In addition, it is observed that using seed morphometric descriptors, the interspecific hybrid shows differences from the three parents, indicating characteristics of the interspecific hybrid determined probably by the hybrid vigor.

Discussion

High-throughput phenotyping methods can facilitate the use of genetic resources by estimating phenotypic traits of importance and identifying accessions of interest for pre-breeding and breeding programs (Nguyen and Norton, 2020). In this work, we explored phenomic descriptors that help discriminate selected *Phaseolus* accessions with their interspecific cross, using components based on physiological descriptors, seed and pod morphometrics and yield components.

Despite being crop relatives of *P. vulgaris*, Acutifolii species (*P. acutifolius* and *P. parvifolius*) have contrasting leaf, seed, and pod phenotypic characteristics (Figure 1). In addition, there are natural differences between domesticated and wild accessions (Mwale et al., 2020). The domestication syndrome of the Phaseolus genus is characterized by a reduction in pod shattering and increase in seed size, being these the most important traits in the adaptation of domesticated populations (Chacón-Sánchez, 2018). This explains the differentiation in our data for seed and pod morphometrics and reveals why they can serve as the most significant traits in quantification of phenomic proportions between the studied hybrid and its parents and under some generalization and verification can be used for a wider spectrum of hybrid evaluations.

Each of the accession's classifications contained several phenomic descriptors that contributed to defining/identifying its uniqueness. In pod and seed morphometry, it is observed that the descriptors with highest contributions (Figures 4A, C), such as seed/pod Area, pod Feret, seed Height, seed MinFeret, seed Minor and seed/pod Major, are descriptors directly related to the organ (seeds or pod) size; while Solidity shows that the pod shape is influenced by its curvature and also by the shape of the seed. *P. parvifolius*, being a wild accession, does not present domestication syndromes (Chacón-Sánchez, 2018). This is evidenced by the pod and seed small size, being the primary discriminating descriptors in classifying the parvifolius accession. Both studied *P. acutifolius* and *P. vulgaris* have larger pod and seed sizes, likely due to the selection pressure of preferable domestication syndromes (Chacón-Sánchez, 2018). The domestication process directly influences pod and seed weights and can increase pod harvest index (PHI) (Rao et al., 2013). Interestingly, *P. acutifolius* generally has smaller seed size (Freytag and Debouck, 2002) but higher PHI compared to *P. vulgaris* (Rao et al., 2013). Higher PHI has strong heritability, is easily measured, and is related to drought resistance and low soil fertility tolerance. The *P. parvifolius* accession also lacks pod shape curvature, allowing simple visual differentiation and classification between accessions.

The physiological descriptors are extremely useful for accession classifications (Figure 5B), although less comparable to morphometric aspects. However, the lower weight of physiological traits is understandable when considering all

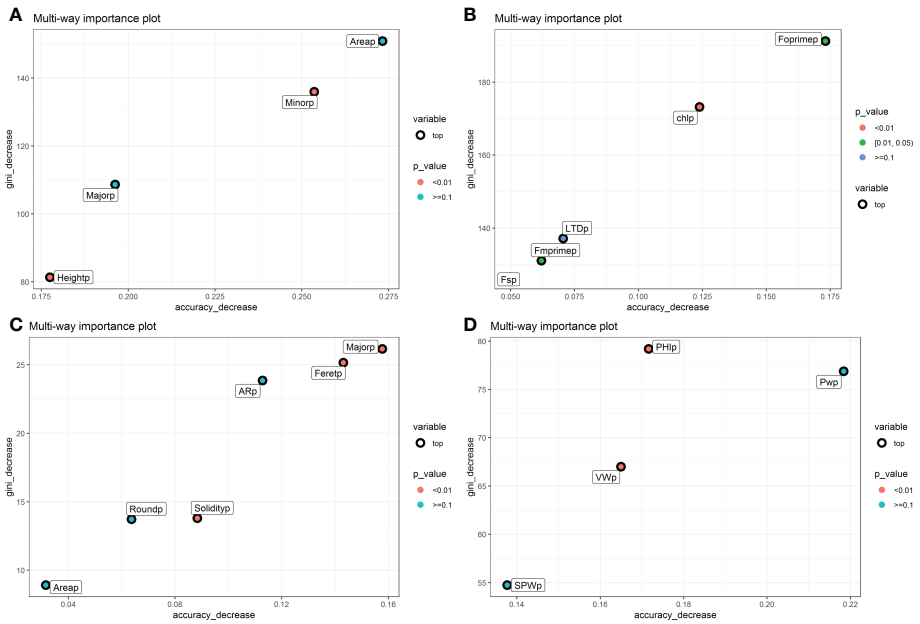


FIGURE 4
Gini index and accuracy decrease for feature selection in random forest for each characterization component in the hybrid. **(A)** Seed morphometric component, **(B)** Physiological component, **(C)** Pod morphometric component and **(D)** Yield component. The weighted phenomic descriptors are observed. The figure shows the phenomic descriptors of major importance in the classification of the parental accessions and the interspecific hybrid.

physiological descriptors (in our case we measured photosynthetic/fluorescence traits and leaf-based data by MultispeQ) as greatly influenced by the environment, with possibly limited heritability and biologically relevant biochemical acclimation thresholds (resistance). The physiological traits inherently hold considerable genetic complexity, considering their crucial role in plant development and survival (Reynolds and Langridge, 2016). Alternatively, it is possible that *P. acutifolius* shows similarly reduced physiological behavior as *P. parvifolius* in the conditions where experiments were done. Although *P. acutifolius* and *P. parvifolius* are two different species (Buhrow, 1983; Schinkel and Gepts, 1989), *P. acutifolius* var. *latifolius* has been reported as an intermediate species between the domesticated *P. acutifolius* and wild *P. parvifolius*. This suggests that both studied accessions may share similar genetic background (Freytag and Debouck, 2002; Muñoz et al., 2006; Blair et al., 2012) and thus some physiological performance as mentioned above.

Regarding the physiological descriptors, it is clear that leaf components as Chl, Fo', Fs', Fm', and LTD (Supplementary Material Figure 4B) present the highest contributions in the differentiation and classification between lines. Both *P. acutifolius* and its wild relative *P. parvifolius*, have similar physiological responses most likely due to the similar ecogeographic distribution of both species, associated with the arid areas of southern USA and northern Mexico (Freytag and Debouck, 2002). Similarly, the above-mentioned descriptors are closely related to photosynthetic efficiency and are recognized - in some scenarios - as indicators of abiotic stress resistance of individual accessions (Sánchez-Reinoso et al., 2019; Guidi et al., 2019). *P. vulgaris* usually exhibits higher sensitivity to drought stress compared to more resistant *P. acutifolius* (Rao et al., 2013; Polania et al., 2016). This can be closely related to the two independent domestication processes of *P. vulgaris* (Chacón et al., 2005), mainly influenced by differences in air/soil humidity and contrasting temperatures between the Mesoamerican and Andean races, presenting differences also in their photosynthetic adaptations (Lynch et al., 1992; González et al., 1995).

The studied interspecific hybrid line INB 47 is a product of interspecific crossings carried out by the CIAT common bean-breeding program. The selection process focused on obtaining adequate seed type, growth habit and yield characteristics from the parent *P. vulgaris*. No surprise then, which the studied interspecific hybrid presented low phenomic proportions with the *P. parvifolius* and *P. acutifolius* parent accessions. This is because agronomically-valued traits likely do not coincide with those two parental accessions. This is probably mainly because physiological traits were selected indirectly (in contrast to agronomically-important descriptors), with no apparent interest/knowledge in/of physiological traits at the time of selection by the breeders (Mejía-Jiménez et al., 1994).

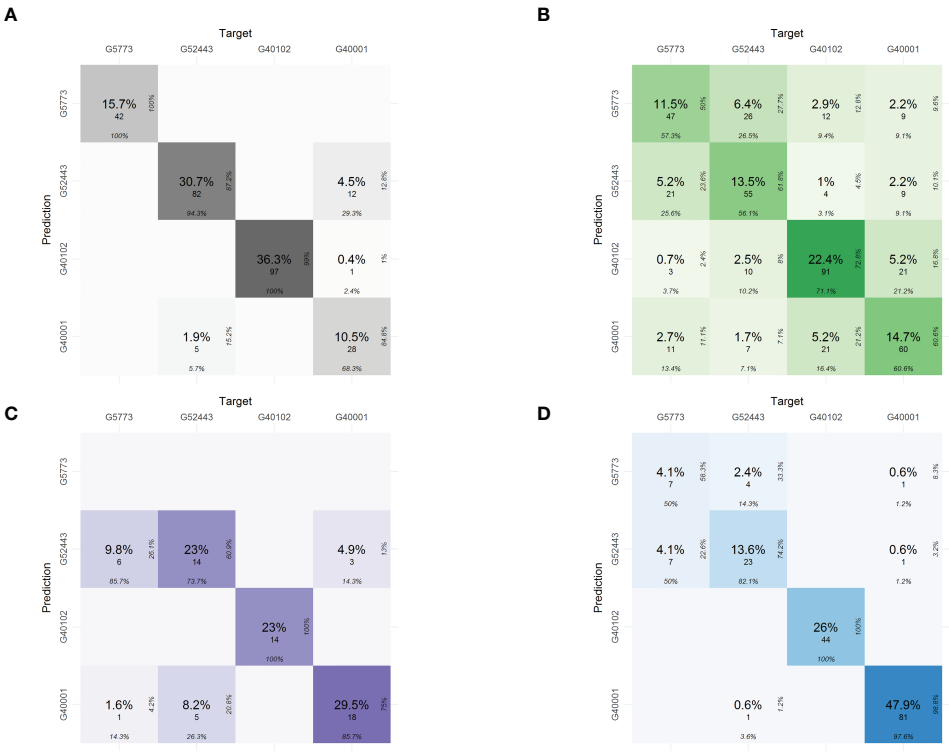


FIGURE 5
Phenomic proportions in the confusion matrices for classification of the interspecific hybrid in each of the characterization components using phenomics descriptors. **(A)** Seed morphometric component, **(B)** Physiological component, **(C)** Pod morphometric component and **(D)** Yield component. The confusion table shows the phenomic proportions of the interspecific hybrid. The phenomic proportions are the predictions of the hybrid with respect to the parental accessions.

Mejía-Jiménez et al. (1994) developed a group of CBC_5 interspecific hybrids with *P. acutifolius*. These authors generated populations with high genetic frequencies of *P. acutifolius*, showing average introgressions of 8% in CBC_5 using amplified fragment length polymorphisms (AFLP). Considering that the interspecific hybridization used in our study employed CBC_5 crossed twice with the parents *P. vulgaris* and *P. parvifolius*, and that it was selected during ten selfing cycles, it is likely that introgression of the *P. acutifolius* has decreased. Nevertheless, the 2.2% of the phenomic proportions of *P. acutifolius* - predicted from the physiological characterization - evidence the successful introgressions from this parental line. In addition, the studied interspecific hybrid also preserves morphological traits similar to *P. acutifolius*, such as the lanceolate leaf shape and acute leaf angle at high light intensities (Figure 1). It could be interesting to evaluate the effect of these morphological traits on the abiotic stress resistance of this or other hybrids (after the methodology generalization).

Moreover, the studied interspecific hybrid keeps some characteristics that can influence the acclimatization process during abiotic stresses (Sánchez-Reinoso et al., 2019). This argues strongly in favor of conserving and characterizing accessions with intermediate phenomic proportions and could allow better understanding and quantifying (based on their GxE base) of the inherited traits and their proportions. It also would support accelerating genetic advances during more effective selection processes based on more newly available data types (semi-automatic remote sensing collection of data of highest interest).

Additionally, in hybrids, phenomic descriptors with the highest distinction powers (discrimination) could allow conserving desired physiological traits of *P. acutifolius* or *P. parvifolius* accessions, without losing key seed characteristics (e.g. size, color or taste) from their *P. vulgaris* parental line. Targeted accession evaluations can be performed by continuous monitoring even during the selection process (starting in already in the F1 generation). This would be based on the suggested machine learning techniques and selected traits of special interest for validating the functional introgressions of desired traits from crop wild relatives.

Our results demonstrate that the use of phenomic descriptors and machine learning analyses offer a very useful alternative for classifying hybrids, by using useful phenotypic and morphometric traits (with some degree of generalization and verification). In reality, breeders focusing on interspecific

10.3389/fpls.2022.1008666

TABLE 1 MANOVA of parents and interspecific hybrid in each of the characterization components.

Characterization component	Accession	Contrast to:	p value	MATS	p - value Resampling
Physiological	*G40001* (*P. acutifolius*)	parvifolius	< 0.001	1016.229	< 0.001
		vulgaris	0.0073		
		hybrid	0.1053		
	G40102 (*P. parvifolius*)	acutifolius	< 0.001		
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G5773 (*P. vulgaris*)	acutifolius	0.0073		
		parvifolius	< 0.001		
		hybrid	0.7523		
Pod morphometrics	*G40001* (*P. acutifolius*)	parvifolius	< 0.001	2538.441	< 0.001
		vulgaris	0.6099		
		hybrid	0.8856		
	G40102 (*P. parvifolius*)	acutifolius	< 0.001		
		vulgaris	0.0004		
		hybrid	< 0.001		
	G5773 (*P. vulgaris*)	acutifolius	0.6099		
		parvifolius	0.0004		
		hybrid	0.8301		
Seed morphometrics	*G40001* (*P. acutifolius*)	parvifolius	< 0.001	26348.12	< 0.001
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G40102 (*P. parvifolius*)	acutifolius	< 0.001		
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G5773 (*P. vulgaris*)	acutifolius	< 0.001		
		parvifolius	< 0.001		
		hybrid	0.6546		
Yield	*G40001* (*P. acutifolius*)	parvifolius	< 0.001	9697.0	< 0.001
		vulgaris	< 0.001		
		hybrid	< 0.001		
	G40102 (*P. parvifolius*)	acutifolius	< 0.001		
		vulgaris	0.0302		
		hybrid	0.015		
	G5773 (*P. vulgaris*)	acutifolius	< 0.001		
		parvifolius	0.0302		
		hybrid	0.9986		

The p-value for each parent and its relationship to the hybrid is observed.

crossings should consider physiological and morphological traits identified in this study as part of an effective screening strategy. This would be especially true where some of these traits were to prove functional in certain environments (willow leaves, leaf angle, growth habit, PHI) or be connected to farmers preferences (seed color and size, pod shape etc.). Breeders would then be able to use other selection criteria apart from the laborious final yield components and seed type characteristics, and thus quickly estimate the introgression efficiency of functional traits in the progenies.

Currently, the CIAT genebank conserves 18 interspecific hybrid lines of *P. vulgaris* x *P. acutifolius* x *P. parvifolius* and 6 interspecific hybrid lines of *P. vulgaris* x *P. acutifolius*, which were selected based on the phenotypic traits conserving characteristics associated with its crop relatives. In addition, the CIAT genebank stores 326 accessions of *P. acutifolius*, including cultivated lines, landraces and wild accessions. However, only a fraction of the whole collection has been studied and characterized for key agronomic and physiological traits, heavily limiting their utilization in pre-breeding or breeding programs (Mwale et al., 2020). This suggests the urgent need to conduct experiments to explore phenomic traits of the *P. acutifolius* collection, including the genetically diverse wild tepary bean accessions as these offer a unique

opportunity to find desirable genes with potential for introducing them into the genetic background of the domesticated tepary bean (Mhlaba et al., 2018; Mwale et al., 2020). Breeders may then be encouraged to start working within the acutifolius group. We believe that selected phenomic descriptors can also help identify suitable "bridge" genotypes for crossings between secondary and tertiary genepools and common beans. The development of phenomic markers (new phenomic proportions recognized as important descriptors) will contribute to germplasm management in genebanks as well (Nguyen et al., 2020). Selected and recognized phenomic descriptors will facilitate the detection of accessions with similar phenomic proportions, determining accessions with high phenomic redundancy, and likely helping germplasm curators even to effectively find duplicate accessions.

Our study demonstrates that selected phenomic descriptors' data processed by a machine learning approach have the potential to discriminate between parental accessions or our studied hybrid. After some generalization (trait verification on different hybrid systems), this methodological approach may help breeders quantify any trait-of-interest introgression directly from different genepools or wild relatives increasing the chances of identifying important consumer target traits in elite common bean lines. After generalization, this methodology also will be able to identify hybrids with hybrid vigor due to the performance of unique phenomic traits. In addition, genome-associated phenomic markers could further contribute to the detection of genes of deep agronomic interest under abiotic and biotic stress conditions (Pasala and Pandey, 2020; Al-Tamimi et al., 2021; Dwivedi et al., 2020; Rassizadeh et al., 2021).

Detailed characterization of CIAT genebank conserved interspecific hybrids or new early breeding materials will likely show new traits with physiological or agronomic potential. In our study, the most contrasting characterization components with the highest precision and stability of the selected *Phaseolus* taxonomy classification are the seed and pod morphometry data.

This study was never intended as the end of classic crop descriptors used by genebanks curators. In reality, classic descriptors will always offer their unique potential. However, some of them can still be rather subjective, are often only qualitative, and require laborious effort to apply them. We have tried to build on the understanding and precision of such classic mostly qualitative descriptors by digitization of some of the crop responses, so as to use a quantitative approach to make some descriptors available to modern breeders and with potential selection power (QTL, GWAS, genomic selection etc.).

In our study we were able to evaluate selected phenomic traits and their ability to become "phenomic markers" and then establish digital descriptors. We also identified machine learning techniques, which allow us to differentiate between studied Phaseolus accessions and determine the similarities or differences of an interspecific hybrid with respect to its parental lines. In our experiment we performed the analysis with random forests, however the strategy can use various machine learning algorithms (Parmley et al., 2019), since the purpose is to determine the most important descriptors that discriminate generally between the parents and its hybrid. There are several algorithms that can have greater accuracy in the classification according to the needs of the researcher and the dimensionality of the phenomic data.

Conclusions

In our work we demonstrate the use of phenomics and machine learning approach as analytical tools in understanding the phenotypic variability of selected Phaseolus accessions and quantifying the crossing effectivity in its related hybrid.

In our study, we quantified the physiological, morphological and yield proportional relatedness of parental lines with its hybrid, finding differences between all groups. Results indicate that the interspecific hybrid preserve intermediate yield characteristics from *P. vulgaris* and *P. acutifolius* parents; although, it has closer phenotypic proportions with *P. vulgaris* (6%). The phenomic proportions method can be a useful tool for the analysis of the closeness of lines/hybrids to their parents even by using traits with clear agronomic potential. However, also physiological data (MultispeQ) showed high potential for lines discrimination, especially towards the studied line of *P. parvifolius*. This complex of traits needs to be further studied and amplified in a wide range of genotypes to verify its value across species, genepools and environments. Our finding provides conclusive evidence that the integration of machine learning, classification algorithms and phenotyping tools promise to automate the precise quantification of phenomics proportion of parents in their hybrids.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

DC: Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work. MU: Drafting the work or revising it critically for important intellectual content. MS: Drafting the work or revising it critically for important intellectual content. JG: Drafting the work or revising it critically for important intellectual content. AC: Analysis, or interpretation of data for the work. PW: Provide approval for publication of the content. All authors contributed to the article and approved the submitted version.

Funding

This work was financed under the Global Challenges Research Fund-Bioinformatics and Biological Resources Project (BB.R01504X/1): Developing a hybrid-bean collection to advance climate-ready bean breeding project, led by the National Institute of Agricultural Botany (NIAB) (93 Lawrence Weaver Road, Cambridge, CB3 0LE, UK).

Acknowledgments

We thank PhD Monica Carvajal and Gustavo Cardona for their help and the space they provided during the collection of morphological and agronomic data. We would also like to thank PhD Daniel Debouck for his deep support and invaluable thoughts. Finally, to the genetic resources program and especially to the regeneration area for the help provided during the agronomic management of the crop. MOU is thankful to GIZ and PIAF Germany for their support. We would like to thank the reviewers who helped to keep the message very clear. Finally we would like to thank Vincent Johnson of the Alliance Science Writing Service for his editorial review.

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.2022.1008666/full#supplementary-material

References

Allendorf, F. W., Leary, R. F., Spruell, P., and Wenburg, J. K. (2001). The problems with hybrids: setting conservation guidelines. *Trends Ecol. Evol.* 16 (11), 613–622. doi: 10.1016/S0169-5347(01)02290-X

Al-Tamimi, N., Oakey, H., Tester, M., and Negrão, S. (2021). "Assessing rice salinity tolerance: From phenomics to association mapping," in *Rice genome engineering and gene editing* (New York, NY: Humana), 339–375.

Araus, J. L., Kefauver, S. C., Vergara-Díaz, O., Gracia-Romero, A., Rezzouk, F. Z., Segarra, J., et al. (2022). Crop phenotyping in a context of global change: What to measure and how to do it. *J. Integr. Plant Biol.* 64 (2), 592–618. doi: 10.1111/jipb.13191

Arnold, M. L. (1997). *Natural hybridization and evolution* (Oxford University Press on Demand: Oxford Series in Ecology and Evolution).

Beebe, S., Ramirez, J., Jarvis, A., Rao, I. M., Mosquera, G., Bueno, J. M., et al. (2011). Genetic improvement of common beans and the challenges of climate change. *Crop Adaptation to Climate Change*, 356–369. doi: 10.1002/9780470960929.ch25

Blair, M. W., Pantoja, W., and Carmenza Muñoz, L. (2012). First use of microsatellite markers in a large collection of cultivated and wild accessions of tepary bean (Phaseolus acutifolius a. Gray). *Theor. Appl. Genet.* 125 (6), 1137–1147. doi: 10.1007/s00122-012-1900-0

Buhrow, R. (1983). The wild beans of southwestern north America. *Desert Plants* 5 (2), 67–88.

Chacón, S. M. I., Pickersgill, B., and Debouck, D. G. (2005). Domestication patterns in common bean (Phaseolus vulgaris l.) and the origin of the mesoamerican and Andean cultivated races. *Theor. Appl. Genet.* 110 (3), 432–444. doi: 10.1007/s00122-004-1842-2

Chacón-Sánchez, M. I. (2018). The domestication syndrome in phaseolus crop plants: A review of two key domestication traits. *Origin Evol. Biodivers.*, 37–59. doi: 10.1007/978-3-319-95954-2_3

Debouck, D. G. (2021). Phaseolus beans (Leguminosae, phaseoleae): A checklist and notes on their taxonomy and ecology. *J. Bot. Res. Inst. Texas* 15 (1), 73–111. doi: 10.17348/jbrit.v15.i1.1052

de Carvalho, M. A., Bebeli, P. J., Bettencourt, E., Costa, G., Dias, S., Dos Santos, T. M., et al. (2013). Cereal landraces genetic resources in worldwide GeneBanks. a review. *Agron. Sustain. Dev.* 33 (1), 177–203. doi: 10.1007/s13593-012-0090-0

Deva, C. R., Urban, M. O., Challinor, A. J., Falloon, P., and Svitákova, L. (2020). Enhanced leaf cooling is a pathway to heat tolerance in common bean. *Front. Plant Sci.* 11, 19. doi: 10.3389/fpls.2020.00019

Dwivedi, S. L., Goldman, I., Ceccarelli, S., and Ortiz, R. (2020). Advanced analytics, phenomics and biotechnology approaches to enhance genetic gains in plant breeding. *Adv. Agron.* 162, 89–142. doi: 10.1016/bs.agron.2020.02.002

Eliceiri, K. W., Berthold, M. R., Goldberg, I. G., Ibá n ez, L., Manjunath, B. S., Martone, M. E., et al. (2012), 697–710.

Feldmann, M. J., Hardigan, M. A., Famula, R. A., Lopez, C. M., Tabb, A., Cole, G. S., et al. (2020). Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry. *GigaScience* 9 (5), giaa030. doi: 10.1093/gigascience/giaa030

Feller, C., Bleiholder, H., Buhr, L., Hack, H., Hess, M., Klose, R., et al. (1995). Phanologische entwicklungsstadien von gemusepflanzen II. fruchtgemuse und hulsenfruchte. *Nachrichtenblatt Des. Deutschen Pflanzenschutzdienstes* 47 (9), 217–232.

Fernández-Calleja, M., Monteagudo, A., Casas, A. M., Boutin, C., Pin, P. A., Morales, F., et al. (2020). Rapid on-site phenotyping *via* field fluorimeter detects differences in photosynthetic performance in a hybrid–parent barley germplasm set. *Sensors* 20 (5), 1486. doi: 10.3390/s20051486

Fernández de Córdoba, F., Gepts, P., and López, M. (1991). Etapas de desarrollo en la planta de fríjol. *Frijol: Investigación y producción*. (Colombia: CIAT) pp: 61–78.

Freytag, G. F., and Debouck, D. G. (2002). Taxonomy, distribution, and ecology of the genus phaseolus (Leguminosae-papilionoideae) in north America, Mexico and central America. *BRIT*. 23, 1–300.

Friedrich, S., and Pauly, M. (2018). MATS: Inference for potentially singular and heteroscedastic MANOVA. *J. Multivariate Anal.* 165, 166–179. doi: 10.1016/j.jmva.2017.12.008

González, A., Lynch, J., Tohme, J. M., Beebe, S. E., and Macchiavelli, R. E. (1995). Characters related to leaf photosynthesis in wild populations and landraces of common bean. *Crop Sci.* 35 (5), 1468–1476. doi: 10.2135/cropsci1995.0011183X003500050034x

Guidi, L., Lo Piccolo, E., and Landi, M. (2019). Chlorophyll fluorescence, photoinhibition and abiotic stress: does it make any difference the fact to be a C3 or C4 species? *Front. Plant Sci.* 10, 174. doi: 10.3389/fpls.2019.00174

Haghighi, K. R., and Ascher, P. D. (1988). Fertile, intermediate hybrids between phaseolus vulgaris and p. acutifolius from congruity backcrossing. *Sexual Plant Reprod.* 1 (1), 51–58. doi: 10.1007/BF00227023

Henao-Rojas, J. C., Rosero-Alpala, M. G., Ortiz-Muñoz, C., Velásquez-Arroyo, C. E., Leon-Rueda, W. A., and Ramírez-Gil, J. G. (2021). Machine learning applications and optimization of clustering methods improve the selection of descriptors in blackberry germplasm banks. *Plants* 10 (2), 247. doi: 10.3390/plants10020247

Huang, L. K., and Wang, M. J. J. (1995). Image thresholding by minimizing the measures of fuzziness. *Pattern Recognition* 28 (1), 41–51. doi: 10.1016/0031-3203(94)E0043-K

Janitza, S., and Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PloS One* 13 (8), e0201904. doi: 10.1371/journal.pone.0201904

Kandanaarachchi, S., and Hyndman, R. J. (2021). Dimension reduction for outlier detection using DOBIN. *J. Comput. Graphical Stat* 30 (1), 204–219. doi: 10.1080/10618600.2020.1807353

Kassambara, A., and Mundt, F. (2020) *Factoextra: extract and visualize the results of multivariate data analyses*. (This R library) Available at: https://CRAN.R-project.org/package=factoextra.

Kennedy, R. E., Yang, Z., Braaten, J., Copass, C., Antonova, N., Jordan, C., et al. (2015). Attribution of disturbance change agent from landsat time-series in support of habitat monitoring in the puget sound region, USA. *Remote Sens. Environ.* 166, 271–285. doi: 10.1016/j.rse.2015.05.005

Kholová, J., Urban, M. O., Cock, J., Arcos, J., Arnaud, E., Aytekin, D., et al. (2021). In pursuit of a better world: crop improvement and the CGIAR. *J. Exp. Bot.* 72 (14), 5158–5179. doi: 10.1093/jxb/erab226

Kholova, J., Hajjarpoor, A., Garin, V., Nelson, W., Diacoumba, M., Messina, C., et al. (2022). The role of crop growth models in crop improvement: integrating phenomics. *Envirotyping Genomic Prediction* doi: 10.19103/AS.2022.0102.13

Kuhlgert, S., Austic, G., Zegarac, R., Osei-Bonsu, I., Hoh, D., Chilvers, M. I., et al. (2016). MultispeQ beta: a tool for large-scale plant phenotyping connected to the open PhotosynQ network. *R. Soc. Open Sci.* 3 (10), 160592. doi: 10.1098/rsos.160592

Kusolwa, P. M., Myers, J. R., Porch, T. G., Trukhina, Y., González-Vélez, A., and Beaver, J. S. (2016). Registration of AO-1012-29-3-3A red kidney bean germplasm line with bean weevil, BCMV, and BCMNV resistance. *J. Plant Registrations* 10 (2), 149–153. doi: 10.3198/jpr2015.10.0064crg

Liaw, M. A., and Wirner, M. (2018). *Package 'randomforest'* (Berkeley, CA, USA: University of California, Berkeley).

Lynch, J., González, A., Tohme, J. M., and García, J. A. (1992). Variation in characters related to leaf photosynthesis in wild bean populations. *Crop Sci.* 32 (3), 633–640. doi: 10.2135/cropsci1992.0011183X003200030012x

Mejía-Jiménez, A., Muñoz, C., Jacobsen, H. J., Roca, W. M., and Singh, S. P. (1994). Interspecific hybridization between common and tepary beans: increased hybrid embryo growth, fertility, and efficiency of hybridization through recurrent and congruity backcrossing. *Theor. Appl. Genet.* 88 (3), 324–331. doi: 10.1007/BF00223640

Mhlaba, Z. B., Mashilo, J., Shimelis, H., Assefa, A. B., and Modi, A. T. (2018). Progress in genetic analysis and breeding of tepary bean (Phaseolus acutifolius a. gray): A review. *Scientia Hortic.* 237, 112–119. doi: 10.1016/j.scienta.2018.04.012

Muñoz, L. C., Duque, M. C., Debouck, D. G., and Blair, M. W. (2006). Taxonomy of tepary bean and wild relatives as determined by amplified fragment length polymorphism (AFLP) markers. *Crop Sci.* 46 (4), 1744–1754. doi: 10.2135/cropsci2005-12-0475

Mwale, S. E., Shimelis, H., Mafongoya, P., and Mashilo, J. (2020). Breeding tepary bean (Phaseolus acutifolius) for drought adaptation: A review. *Plant Breed.* 139 (5), 821–833. doi: 10.1111/pbr.12806

Nankar, A. N., Tringovska, I., Grozeva, S., Ganeva, D., and Kostova, D. (2020). Tomato phenotypic diversity determined by combined approaches of conventional and high-throughput tomato analyzer phenotyping. *Plants* 9 (2), 197. doi: 10.3390/plants9020197

Nguyen, G. N., and Norton, S. L. (2020). Genebank phenomics: A strategic approach to enhance value and utilization of crop germplasm. *Plants* 9 (7), 817. doi: 10.3390/plants9070817

Noriega, I. L., Halewood, M., Abberton, M., Amri, A., Angarawai, I. I., Anglin, N., et al. (2019). CGIAR operations under the plant treaty framework. *Crop Sci.* 59 (3), 819–832. doi: 10.2135/cropsci2018.08.0526

Paluszynska, A. (2017). *Structure mining and knowledge extraction from random forest with applications to the cancer genome atlas project* (University of Warsaw: Master's thesis in MATHEMATICS in the eld of APPLIED MATHEMATICS).

Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S., and Singh, A. K. (2019). Machine learning approach for prescriptive plant breeding. *Sci. Rep.* 9 (1), 1–12. doi: 10.1038/s41598-019-53451-4

Pasala, R., and Pandey, B. B. (2020). Plant phenomics: High-throughput technology for accelerating genomics. *J. Biosci.* 45 (1), 1–6. doi: 10.1007/s12038-020-00083-w

Polania, J., Rao, I. M., Cajiao, C., Rivera, M., Raatz, B., and Beebe, S. (2016). Physiological traits associated with drought resistance in Andean and mesoamerican genotypes of common bean (Phaseolus vulgaris l.). *Euphytica* 210 (1), 17–29. doi: 10.1007/s10681-016-1691-5

Porch, T. G., Beaver, J. S., and Brick, M. A. (2013). Registration of tepary germplasm with multiple-stress tolerance, TARS-tep 22 and TARS-tep 32. *J. Plant Registrations* 7 (3), 358–364. doi: 10.3198/jpr2012.10.0047crg

Reynolds, M., and Langridge, P. (2016). "Physiological breeding." *Curr. Opin. Plant Biol.* 31 (2016), 162–171.

Rao, I., Beebe, S., Polania, J., Ricaurte, J., Cajiao, C., Garcia, R., et al. (2013). Can tepary bean be a model for improvement of drought resistance in common bean? *Afr. Crop Sci. J.* 21 (4), 265–281. doi: 10.4314/ACSJ.V21I4.

Rassizadeh, L., Cervero, R., Flors, V., and Gamir, J. (2021). Extracellular DNA as an elicitor of broad-spectrum resistance in arabidopsis thaliana. *Plant Sci: An International J. Experimental Plant Biol.*312, 111036. doi: 10.1016/j.plantsci.2021.111036

Rosero, A., Pérez, J. L., Rosero, D., Burgos-Paz, W., Martínez, R., Morelo, J., et al. (2019). Morphometric and colorimetric tools to dissect morphological diversity: An application in sweet potato [Ipomoea batatas (L.) lam.]. *Genet. Resour. Crop Evol.* 66 (6), 1257–1278. doi: 10.1007/s10722-019-00781-x

Salazar, D. E., Santos, L. G., Wenzl, P., and Hay, F. R. (2020). Effect of dry heat on seed germination of desmodium and stylosanthes species. *Seed Sci. Technol.* 48 (3), 419–437. doi: 10.15258/sst.2020.48.3.11

Sánchez-Reinoso, A. D., Ligarreto-Moreno, G. A., and Restrepo-Díaz, H. (2019). Chlorophyll α fluorescence parameters as an indicator to identify drought susceptibility in common bush bean. *Agronomy* 9 (9), 526. doi: 10.3390/agronomy9090526

Schinkel, C., and Gepts, P. (1989). Allozyme variability in the tepary bean, phaseolus acutifolius a. Gray. *Plant Breed.* 102 (3), 182–195. doi: 10.1111/j.1439-0523.1989.tb00336.x

Schlautman, B., Diaz-Garcia, L., and Barriball, S. (2020). Morphometric approaches to promote the use of exotic germplasm for improved food security and resilience to climate change: a kura clover example. *Plant Sci.* 290, 110319. doi: 10.1016/j.plantsci.2019.110319

Singh, S. P., Debouck, D. G., and Roca, W. (1998). *Interspecific hybridization between phaseolus vulgaris l. and p. parvifolius freytag.* Annual Report (USA). 41, 7–8.

Singh, A., Ganapathysubramanian, B., Singh, A. K., and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* 21 (2), 110–124. doi: 10.1016/j.tplants.2015.10.015

Singh, S. P., and Munoz, C. G. (1999). Resistance to common bacterial blight among phaseolus species and common bean improvement. *Crop Sci.* 39 (1), 80–89. doi: 10.2135/cropsci1999.0011183X003900010013x

Soltis, P. S., Nelson, G., Zare, A., and Meineke, E. K. (2020). Plants meet machines: Prospects in machine learning for plant biology. *Appl. Plant Sci.* 8 (6), 1–6. doi: 10.1002/aps3.11371

Tadesse, W., Sanchez-Garcia, M., Assefa, S. G., Amri, A., Bishaw, Z., Ogbonnaya, F. C., et al. (2019). Genetic gains in wheat breeding and its role in feeding the world. *Crop Breed. Genet. Genom* 1, e190005.

Wang, S. M., and Zhang, Z. W. (2011). The state of the world's plant genetic resources for food and agriculture. *J. Plant Genet. Resour.* 12 (3), 325–338.

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fpls.2022.1009244

OPEN ACCESS

EDITED BY
Jinyoung Y. Barnaby,
United States Department of Agriculture (USDA), United States

REVIEWED BY
Sivakumar Sukumaran,
The University of Queensland, Australia
Shuwei Liu,
Shandong University (Qingdao), China

*CORRESPONDENCE
Behailu Mulugeta
behailu.mulugeta@slu.se

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

RECEIVED 01 August 2022
ACCEPTED 20 December 2022
PUBLISHED 26 January 2023

CITATION
Mulugeta B, Tesfaye K, Ortiz R, Johansson E, Hailesilassie T, Hammenhag C, Hailu F and Geleta M (2023) Marker-trait association analyses revealed major novel QTLs for grain yield and related traits in durum wheat.
Front. Plant Sci. 13:1009244.
doi: 10.3389/fpls.2022.1009244

Marker-trait association analyses revealed major novel QTLs for grain yield and related traits in durum wheat

Behailu Mulugeta[1,2,3]*, Kassahun Tesfaye[1,4], Rodomiro Ortiz[2], Eva Johansson[2], Teklehaimanot Hailesilassie[1], Cecilia Hammenhag[2], Faris Hailu[5] and Mulatu Geleta[2]

[1]Institute of Biotechnology, Addis Ababa University, Addis Ababa, Ethiopia, [2]Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden, [3]Sinana Agricultural Research Center, Oromia Agricultural Research Institute, Bale-Robe, Ethiopia, [4]Director General, Bio and Emerging Technology Institute (BETin), Addis Ababa, Ethiopia, [5]Department of Biology and Biotechnology, Wollo University, Dessie, Ethiopia

The growing global demand for wheat for food is rising due to the influence of population growth and climate change. The dissection of complex traits by employing a genome-wide association study (GWAS) allows the identification of DNA markers associated with complex traits to improve the productivity of crops. We used GWAS with 10,045 single nucleotide polymorphism (SNP) markers to search for genomic regions associated with grain yield and related traits based on diverse panels of Ethiopian durum wheat. In Ethiopia, multi-environment trials of the genotypes were carried out at five locations. The genotyping was conducted using the 25k Illumina Wheat SNP array to explore population structure, linkage disequilibrium (LD), and marker-trait associations (MTAs). For GWAS, the multi-locus Fixed and Random Model Circulating Probability Unification (FarmCPU) model was applied. Broad-sense heritability estimates were high, ranging from 0.63 (for grain yield) to 0.97 (for thousand-kernel weight). The population structure based on principal component analysis, and model-based cluster analysis revealed two genetically distinct clusters with limited admixtures. The LD among SNPs declined within the range of 2.02–10.04 Mbp with an average of 4.28 Mbp. The GWAS scan based on the mean performance of the genotypes across the environments identified 44 significant MTAs across the chromosomes. Twenty-six of these MTAs are novel, whereas the remaining 18 were previously reported and confirmed in this study. We also identified candidate genes for the novel loci potentially regulating the traits. Hence, this study highlights the significance of the Ethiopian durum wheat gene pool for improving durum wheat globally. Furthermore, a breeding strategy focusing on accumulating favorable alleles at these loci could improve durum wheat production in the East African highlands and elsewhere.

KEYWORDS

candidate gene, durum wheat, GWAS, linkage disequilibrium, population structure, QTL

1 Introduction

Durum wheat (*Triticum turgidum*, L. var. *durum* Desf.) is a staple cereal crop produced to make pasta, bread, and other traditional food items (Sall et al., 2019; Ceglar et al., 2021). Durum wheat accounts for approximately 8% of global wheat production (Sall et al., 2019), and most of it (75%) is produced in the Mediterranean region (Xynias et al., 2020). The world's largest producers are Turkey and Canada, while Ethiopia is the largest producer in Sub-Saharan Africa (SSA). Durum wheat was domesticated around 10,000 years ago in the Fertile Crescent (Özkan et al., 2002; Soriano et al., 2018; Ceglar et al., 2021) and has since then been a vital source of energy, minerals, and bioactive compounds in human nutrition (Johansson et al., 2020a). The durum wheat is an amphidiploid species containing an AABB genome, and its genome size is nearly 12 Gb (Maccaferri et al., 2019).

The current development of advanced DNA sequencing methods, functional genomic tools, and availability of different DNA chip technology has highly facilitated the genetic dissection of multi-genic traits of food crops (Collard and Mackill, 2008; Al-Khayri et al., 2016; Geleta and Ortiz, 2016). Association mapping (AM) has been widely used to dissect the genetic architecture behind traits like grain yield, host plant resistance to pathogens, drought and salinity tolerance, phenology, and quality traits (Maccaferri et al., 2010; Tuberosa, 2012; Canè et al., 2014; Turki et al., 2015; Giraldo et al., 2016; Mengistu et al., 2016; Kidane et al., 2019; Mérida-García et al., 2019; Mérida-García et al., 2020). Moreover, genome-wide association studies (GWAS) have been successfully used to map genetic loci and dissect the genomic regions underlying several vital traits in important food crops, such as barley (Bellucci et al., 2017; Borrego-Benjumea et al., 2021), and bread wheat (Li et al., 2019; Gao et al., 2021; Mekonnen et al., 2021).

In wheat, GWAS has been successfully applied to identify and dissect QTL associated with grain yield (Li et al., 2019; Gao et al., 2021), host plant resistance to pathogens (Alemu et al., 2021a; Alemui et al., 2021; Mekonnen et al., 2021), drought tolerance (Bhatta et al., 2018; Mathew et al., 2019), root architecture (Alemu et al., 2021b), phenology (Mekonnen et al., 2021), adaptation to salinity (Quamruzzaman et al., 2021), and end-use quality traits (Chen et al., 2019; Talini et al., 2020). However, in durum wheat, limited GWAS results have been reported across traits of interest, although some results are present for grain yield (Wang et al., 2019; Anuarbek et al., 2020), host plant resistance to pathogens (Liu et al., 2017b; Aoun et al., 2021), drought tolerance (Wang et al., 2019), root system architecture (Maccaferri et al., 2016; Alemu et al., 2021b), osmotic adjustment (Condorell et al., 2022), and phenology and quality traits (Fiedler et al., 2017). Furthermore, GWAS results reported on Ethiopian durum wheat cultivars and landraces are insufficient. Increased genomic research is needed to improve durum wheat production in Ethiopia by utilizing genomic-assisted breeding approaches.

Wheat landraces can be seen as an essential germplasm resource, with the potential to be utilized as a reservoir of crop diversity that harbors significant novel loci associated with agronomic, phenological, and end-use quality traits (Johansson et al., 2021). Landraces and their wild relatives have served as sources of valuable genes to improve modern cultivars for adaptation to diverse environments, grain yield, end-use quality, host plant resistance to the pathogen, and abiotic stress tolerance (Maccaferri et al., 2019; Johansson et al., 2020b; Sansaloni et al., 2020). Several reports revealed that Ethiopian durum wheat has high genetic diversity to be explored in the search for essential novel and valuable genes for improvements of traits such as grain yield, nutritional quality, host plant resistance to pathogens, and drought tolerance (Mengistu et al., 2016; Kabbaj et al., 2017; Kidane et al., 2019; Alemu et al., 2020a). Hence, understanding the genetic basis of these important traits using recent genomic-based research will facilitate the use of Ethiopian germplasm in an improvement program to maintain a food-secure future in the region.

This study aimed to use GWAS to define genomic regions in Ethiopian durum wheat associated with grain yield and related traits. Furthermore, population structure and linkage disequilibrium were evaluated for precise identification of the genetic basis of valuable genomic regions associated with grain yield and important agronomic traits.

2 Materials and methods

2.1 Germplasm

The present study used 420 Ethiopian durum wheat landraces and cultivars. To accommodate the extensive diversity of the Ethiopian durum wheat gene pool, 385 landraces were selected from different geographical regions of Ethiopia, while 35 were crossbred cultivars. Supplementary Table 1 provides information on these landraces and cultivars. For simplicity, the landraces and cultivars will be designated as genotypes hereafter.

2.2 Description of test environment

The performance of the genotypes was evaluated across five locations in Ethiopia, namely, Akaki (AK; 09°53' 48" N/39°49' 16" E), Chefe Donsa (CD; 08°58' 57" N/39°09' 13" E), Holeta (HO; 09°01' 15" N/38°28′ 26" E), Kulumsa (KU; 08°01' 11" N/39°09' 37" E) and Sinana (SN; 07°06′ 58" N/40°13′ 38" E)

during the 2019 - 2020 main crop-growing season. The testing locations represent the country's major and most suitable durum wheat growing environments. The soil texture of each site is characterized as heavy clay for Akaki and Chefe Donsa and clay for Holeta, Kulumsa, and Sinana. The test sites are classified into ME (Mega environment)2:SW(Spring Wheat) high rainfall areas that receive more than 500 mm of rainfall during the crop growing cycle as defined by CIMMYT's Wheat Breeding Program (Rajaram et al., 1994). Among the five test sites, Sinana, Kulumsa, and Holeta have been used by CIMMYT's wheat breeding program targeting high potential environments in the highlands of East Africa. The Agro-ecology at the Akaki and Chefe Donsa sites are also similar to those at the other three test sites and are considered high-potential sites. During the crop-growing season, the mean monthly maximum and minimum temperature of the Sinana site ranged from 20.8°C to 23.9°C and 8.4°C to 9.2°C, respectively, with total rainfall of 810 mm (Supplementary Table 2). The Holeta site received a total rainfall of 852 mm, with the mean annual minimum and maximum temperature of 10°C and 24°C, respectively (Supplementary Table 2). The Chefe Donsa site received mean monthly minimum and maximum temperatures ranging from 9.4–11.8°C and 20.3–24.2°C, with a total rainfall of 870.5 mm. Whereas, the Akaki site received mean monthly minimum and maximum temperatures ranging from 9.89–13.82°C and 24.85–26.91°C, respectively, with a total rainfall amount of 711.5 mm. Kulumsa site received a total rainfall of 700 mm, with a mean monthly minimum temperature of 11°C and a mean monthly maximum temperature of 23°C.

2.3 Field experimental design

The experiment was laid out using an alpha lattice design with two replications containing 21 incomplete blocks with a block size of 20, according to Patterson and Williams (1976). The landraces and cultivars were randomly assigned and planted on a plot size of 1 m^2 with 2.5 m x 0.4 m (two rows with 20 cm spacing). The space between the plots was 20 cm. A seed rate of 150 kg ha^{-1} and fertilizer rate of 50 kg N ha^{-1} and 100 kg of P_2O_5 ha^{-1} was applied to each plot. In order to maintain genotype uniformity (since the genotypes were mostly landraces with possible seed admixture), the genotypes were grown on different plots for two consecutive crop growing seasons (2017-2018) at Sinana agricultural research center, and individual plants that appeared to differ in any of the clearly visible phenotypic traits were removed.

2.4 Evaluation of phenotypic traits

In this study, phenotyping was conducted by applying the previously described methodology for evaluating wheat genetic resources (IBPGR, 1985). The traits measured were phenology (days to heading, days to physiological maturity, and grain filling period), plant architecture (plant height, spike length, and number of effective tillers per plant), grain yield, and grain yield-related traits (number of spikelets per spike, and thousand kernel weight).

2.5 Statistical analysis of the phenotypic data

Before further analysis, data were evaluated by the Shapiro–Wilk test to assess if they fit into the normal distribution. Furthermore, based on the results from the normality test, the homogeneity test was performed for the scored data in the experiment as described in Levene (1960). The R statistical software (R Development Core team, 2021) was used for computing descriptive statistics (mean, range, standard deviation), coefficient of variation, analysis of variance (ANOVA), correlation among traits, and broad-sense heritability. The linear mixed model (LMM) fitted by the Restricted/Estimated Maximum Likelihood method [REML, Corbeil and Searle (1976)] in R package "lme4" (Bates et al., 2015) was used to estimate the variance components of scored traits. To perform ANOVA for each test environment, the genotypes and blocks were considered fixed and random effects, respectively. The response of the i^{th} genotype in the j^{th} incomplete block with the l^{th} replication of each environment for a particular trait was described as:

$$Y_{ijl}=\mu + \tau_i+\beta_j+\gamma_{l(j)}+\xi_{ijl} \quad (1)$$

where Y_{ijl} is the phenotypic response of the i^{th} genotype in l^{th} incomplete block within j^{th} replication, μ is the overall mean, τ_i is the fixed effect of genotype i, β_j is the random effect of the j^{th} replicate, γ_l is the random effect of the j^{th} incomplete block nested in the l^{th} replication, and ξ_{ijl} is the residual error.

The combined ANOVA across environments inference was computed for all the response variables as follows:

$$Y_{ijlk} = \boldsymbol{\mu} + \tau_i + \beta_l + \gamma_j + E_k + \beta\gamma_{lj} + \gamma E_{jk} + \tau E_{ik} + \xi_{ijlk} \quad (2)$$

where Y_{ijlk} is the observed phenotypic trait for i^{th} genotype in l^{th} incomplete block within j^{th} replication at the k^{th} environment, μ is the overall mean, τ_i is the fixed effect of genotype i, β_l is the random effect of j^{th} replication, γ_j is the random effect of the l^{th} incomplete block within j^{th} replication, E_k is the random effect of environment k, $\beta\gamma_{lj}$ is random effect of incomplete block l nested in replication j, γE_{jk} is random effect of replication j in test environment k, τE_{ik} is random effect of interaction between genotype i and environment k, and ξ_{ijlk} is a random residual effect. For the sake of simplicity, we assumed that all the underlying random effects residuals are normally distributed with zero mean and are independent homoscedastic.

The best linear unbiased estimates (BLUEs) of measured traits for each genotype from each environment were obtained using META R software (Alvarado et al., 2020). The estimated means of BLUEs was used to compute the Pearson correlation coefficient (r) by the "cor. test" function in the R (R Development Core team, 2021) and GWAS analysis. The estimates of broad-sense heritability (H2) were computed from pooled ANOVA across environments (Gonçalves-Vidigal et al., 2008) as:

$$H^2 = \frac{\sigma^2_g}{[\sigma^2_g + (\sigma^2_{gl}/l) + (\sigma^2_e/lr)]} \quad (3)$$

where σ^2_g is genotypic variance, σ^2_{gl} is genotype by environment interaction variance, σ^2_e is environmental variance, **l** is the number of environments, and **r** is the number of replications.

2.6 DNA extraction, genotyping, and filtering of SNP markers

A single spike representing each genotype was collected during field phenotyping for genotyping. Five healthy seeds from each spike were taken to represent each genotype and were planted in 3 L pots in a greenhouse at the Swedish University of Agricultural Science (SLU), Alnarp, Sweden. A total of ten 6 mm leaf discs sampled from five two-week-old seedlings of each genotype were collected in each well of a 96-deep well plate and freeze-dried using the CoolSafe ScanVAC Freeze Dryer according to the instructions provided by Trait Genetics. The freeze-dried samples in 96-well deep well plates were sent to TraitGenetics (GmbH, Gatersleben, Germany) for DNA extraction and subsequent genotyping. A standard cetyltrimethylammonium bromide (CTAB) protocol was used to extract DNA from the leaf samples in TraitGenetics' lab. The 420 genotypes were genotyped using an Illumina Infinium 25k wheat single nucleotide polymorphism (SNP) array following the manufacturer´s protocol. The details of the SNP array can be found at https://www.traitgenetics.com/index.php/service-products. Based on a specific durum wheat cluster file developed by TraitGenetics that differentiates durum wheat from bread wheat, markers accurately scored for the A and B genomes were recorded.

Several criteria were used to filter the genotypic data obtained before further analysis. TASSEL 5.2.80 software (Bradbury et al., 2007) was used to remove SNP loci with missing data above 5% or with minor allele frequency (MAF) below 5% (including monomorphic loci). Further filtering of the remaining SNP loci was conducted based on the level of observed heterozygosity (Ho), and loci with Ho greater than 0.01 were excluded. These filtering steps resulted in 10,045 SNPs that were used for data analyses. The evaluation of these SNP loci showed that each of the 420 samples had less than 1% missing data, and hence no genotype was excluded from the data analyses.

2.7 Population structure and linkage disequilibrium (LD) analysis

The number of subgroups among the 420 genotypes was inferred by principal component analysis (PCA) and model-based clustering methods, which were computed by Genome Association and Integrated Prediction Tool (GAPIT) 3.0 (Wang and Zhang, 2021) and STRUCTURE 2.3.4. software (Pritchard et al., 2000; Falush et al., 2007), respectively. A Bayesian approach (MCMC: Markov Chain Monte Carlo) that assumes an ancestry model of ADMIXTURE and correlated allele frequencies among the subgroups was used for model-based cluster analysis. The length of the burn-in period was adjusted to 50,000, followed by 100,000 MCMC iterations for subgroups (K) ranging from one to ten. Ten independent runs were carried out for each K. The STRUCTURE results were visualized using STRUCTURE Harvester (Earl and vonHoldt, 2012). The number of best K was inferred using the delta K method described in Evanno et al. (2005). The optimum K value bar plot was drawn based on CLUMPAK online software (Kopelman et al., 2015).

Information on the pattern of linkage disequilibrium (LD) within a genetic material of interest is necessary to determine the marker density required for a genome-wide scan (Siol et al., 2017). Accordingly, LD was computed using TASSEL version 5.2.8 (Bradbury et al., 2007). The pairwise LD (squared allele frequency, r^2) for pairs of SNP markers was computed according to Weir (1997). The intersection of the fitted curve with the cut-off threshold was considered the mean r^2 value for each chromosome (Breseghello and Sorrells, 2006b; Liu et al., 2017c). The mean r^2 value of each chromosome was computed and plotted against the chromosome's physical distance. The physical distance at which the r^2 value dropped to half its average maximum value was considered the LD decay rate (Huang et al., 2010). The $r^2 = 0.3$ ($p<0.01$) was considered as a cut-off point to represent a limit of QTL between pairs of markers as indicated in previous studies for Ethiopian durum wheat panels (Liu et al., 2017c; Alemu et al., 2021b).

2.8 Identification of marker trait association

GWAS was conducted using best linear unbiased estimates (BLUEs) for nine phenotypic traits and 10,045 SNP markers to identify marker-trait association (MTA). The BLUEs for grain yield, spike length, and grain-filling period were calculated by considering days to heading (DTH) as a covariate in order to control the effect of heading time, as suggested in previous studies (Sabadin et al., 2012; Tuberosa, 2012). The analysis was

performed by employing a multi-locus-based method, fixed and random model Circulating Probability Unification [FarmCPU, Liu et al. (2016)] model selection algorithm implemented in GAPIT 3 R package (Lipka et al., 2012; Tang et al., 2016; Wang and Zhang, 2021). The FarmCPU model algorithm eliminates potential confounding factors by employing the fixed and random effect models iteratively. This was done to overcome the overfitting model influences of the stepwise regression and to control spurious MTA caused by population structure and family relatedness. GAPIT 3 was also used to visualize the Manhattan and Quantile-quantile (QQ)-plots. The QQ-plot fits the model to account for the population structure.

The stringent false-positive discovery rate [FDR, $p< 0.01$ (Benjamini and Hochberg, 1995)] and Bonferroni-corrected threshold of (–log10 (0.05/n) = 5.30 was used, where n is the total numbers of SNPs) to declare a significant MTA between a marker and phenotypic trait. All MTAs above the threshold levels were rated as significant. The percentage of phenotypic variance explained (PVE) by individual MTA (Garcia et al., 2019) and a marker-based VanRaden kinship (K) matrix (VanRaden, 2008) for the genotypes of interest was also generated in R/GAPIT 3. It was assumed that an identified QTL is stable in the genomic region when significant MTA has appeared in two or more test locations, and the additive effects were concordant.

2.9 Identification of putative novel MTAs and associated candidate genes

The novelty of significant MTAs and their potential associated genes were determined by comparative analyses with previously published reports using different *Triticum* databases such as GrainGene, T3/wheat, and Wheat URGI (Alaux et al., 2018). The lists of different genes and functions were downloaded from the NCBI database (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/900/231/445/GCA_900231445.1_Svevo.v1/) to identify genes related to significant MTAs. The nucleotide position extending from 1-5 cM up and downstream from the SNP position was used for searching the potential candidate genes, as previously reported for wheat (Breseghello and Sorrells, 2006a). The genes associated with the significant MTAs were obtained from the durum wheat (*Triticum turgidum* (Svevo.v1) reference genome) (Maccaferri et al., 2019).

3 Results

3.1 Phenotypic mean performance of genotypes

Descriptive statistics, frequency distribution, and boxplots clearly showed a wide range of variation for all the traits evaluated (Table 1 and Figure 1). The mean number of days for days to heading, days to physiological maturity, and grain filling period over the combined environments were 72.6, 136, and 63.4, respectively (Table 1). The mean grain yield was 6.7 t ha^{-1}, while the mean thousand kernel weight was 40.9 g. On average, 17.5 spikelets per spike were recorded across the environments (Table 1). The highest mean grain yield (8.4 t ha^{-1}) was observed at Chefe Donsa, followed by Sinana (8.2 t ha^{-1}), whereas the lowest mean grain yield was recorded at Holeta (4.4 t ha^{-1}) (Table 1). The highest mean performance of the genotypes for the thousand kernel weight (42.7 g) was attained at Chefe Donsa, whereas the lowest was found at Akaki (35.2 g). The pooled ANOVA over test environments indicated a significant ($p<0.01$) impact of genotype, environment, and genotype by environment interactions on all traits evaluated (Supplementary Table 3). Furthermore, significant effects of replications and blocks were noted for traits, most likely due to variation within the field.

3.2 Variance components estimation, broad-sense heritability, and relationship between traits

The estimates of genotypic variance (σ_g^2) and genotypic coefficient of variation (GCV) for the thousand kernel weight (TKW) and grain-filling period (GFP) were high. The lowest σ_g^2 and GCV were obtained for grain yield (GYD) and the number of effective tillers (NET; Supplementary Table 4). The highest values of variance due to genotype by environment interaction (σ_{gxe}^2) and variance due to environments (σ_{gxe}^2) were recorded for plant height (PHT). In contrast, the number of effective tillers per plant (NET) showed the lowest values of both variances. The phenotypic coefficient of variation (PCV) ranged from 24.1 for days to maturity (DTM) to 137.5 for TKW. Most of the phenotypic traits evaluated in the present study showed high heritability (Supplementary Table 4). The highest broad sense heritability values were recorded for TKW (H^2 = 0.97) and GFP (H^2 = 0.98), indicating that these traits are highly heritable.

The Pearson correlation coefficients computed based on the BLUE mean values were positively significant ($p< 0.01$) for DTH with DTM, SPP, SPL, PHT and NET, for SPP with SPL, NET and GYD, and for SPL with PHT and NET (Figure 2). GYD was positively correlated with SPP (r = 0.20), and TKW (r = 0.24). Nevertheless, GYD had a negative correlation with DTH (r = -0.22) and PHT (r = -0.38) (Figure 2). DTH had a negative correlation with GFP and a positive correlation with DTM and SPP.

3.3 SNP markers distribution and density

In total, the 420 genotypes were genotyped with 24,145 SNP markers. Filtering of the genotypic data based on the number of

TABLE 1 Descriptive statistics for days to heading (DTH, in days), days to physiological maturity (DTM, in days), grain filling period (GFP, in days), plant height (PHT, in cm), spike length (SPL, in cm), grain yield (GYD, in t ha^{-1}), thousand kernel weight (TKW, in g), and the number of spikelets per spike (SPP, in counts) of 420 durum wheat genotypes grown in five test sites (ENV) in Ethiopia.

Traits	ENV	Mean	Median	Range	SE [z]	Traits	ENV	Mean	Median	Range	SE
DTH	Akaki	77.3	78	59–86	0.19	SPL	Akaki	6.24	6	4–12	0.04
	Chefe Donsa	74.8	76	68–84	0.14		Chefe Donsa	7.69	8	5–14	0.05
	Holeta	74.9	76	64–84	0.14		Holeta	8.16	8	5–13	0.06
	Kulumsa	68.0	68	59–77	0.13		Kulumsa	8.1	8	5–12	0.05
	Sinana	64.1	68	59–76	0.13		Sinana	9.48	10	4–14	0.07
	Pooled ENV	72.6	73	59–86	0.08		Pooled ENV	7.94	8	4–14	0.03
DTM	Akaki	138.5	137	133–153	0.12	GYD	Akaki	4.9	4.9	1.2–9.9	0.05
	Chefe Donsa	137.3	138	131–150	0.09		Chefe Donsa	8.4	8.5	1.6–13	0.06
	Holeta	132.6	133	127–147	0.09		Holeta	4.36	4.3	2.7–11	0.03
	Kulumsa	133.6	134	127–145	0.09		Kulumsa	7.78	7.8	1.8–13.5	0.06
	Sinana	138.1	138	130–150	0.08		Sinana	8.2	8.1	2.9–14	0.07
	Pooled ENV	136.0	136	127–153	0.06		Pooled ENV	6.74	6.7	1.2–14	0.03
GFP	Akaki	61.3	60	48–81	0.17	TKW	Akaki	35.25	35	20–51	0.16
	Chefe Donsa	62.5	62	51–77	0.13		Chefe Donsa	42.69	42	27–60	0.18
	Holeta	57.7	57	48–75	0.15		Holeta	42.21	42	29–60	0.17
	Kulumsa	65.6	65	53–80	0.16		Kulumsa	42.25	42	26–66	0.20
	Sinana	70.0	70	58–85	0.14		Sinana	42.05	41.5	21–57	0.18
	Pooled ENV	63.4	63	48–85	0.09		Pooled ENV	40.89	41	20–66	0.09
PHT	Akaki	69.5	69	43–111	0.31	SPP	Akaki	15.84	16	9–23	0.07
	Chefe Donsa	100.0	100	70–133	0.33		Chefe Donsa	18.14	18	13–23	0.06
	Holeta	93.4	94	57–129	0.40		Holeta	14.58	15	10–23	0.05
	Kulumsa	101.2	102	58–140	0.37		Kulumsa	18.14	18	13–24	0.06
	Sinana	117.0	117	65–156	0.40		Sinana	20.88	21	14–26	0.06
	Pooled ENV	96.3	98	43–156	0.28		Pooled ENV	17.52	18	9–26	0.04

[z]SE, Standard error.

missing values, observed heterozygosity, and minor allele frequency resulted in 10,045 high quality polymorphic SNPs used for data analyses. Of these 10,045 SNPs, 4,807 (48%) and 5,238 (52%) were distributed on the A and B genomes, respectively (Table 2). The number of these SNPs per chromosome with regard to the two genomes ranged from 415 on chromosome 4B to 917 on chromosome 5B (Table 2).

The marker density was 1.01, 0.96, and 0.98 Mbp per marker for the A, B, and whole genomes, respectively. The SNP markers used in this study covered a total size of 9.86 Gbps, with chromosomes 1A and 2B having the smallest (584.2 Mbp) and largest (789.4 Mbp) regions (Table 2 and Figure 3).

3.4 Linkage disequilibrium

Among all possible pairs of SNPs on each chromosome, 490,775 pairs were found in LD (Table 3). Of these, 97,386 (19.8%) were found to be significant marker pairs with $r^2 \geq 0.3$ ($p < 0.01$; Table 3), which were therefore used to assess the MTAs. The significant marker pairs on each chromosome accounted for 12.1% ($r^2 = 0.12$ for chromosome 7A) to 26.2% ($r^2 = 0.211$ for chromosome 3B) of all marker pairs on the corresponding chromosomes (Table 3). The sudden LD decay among SNP pairs occurred within the range of 2.02–10.04 Mbp with an average of 4.26 Mbp (Supplementary Table 5;

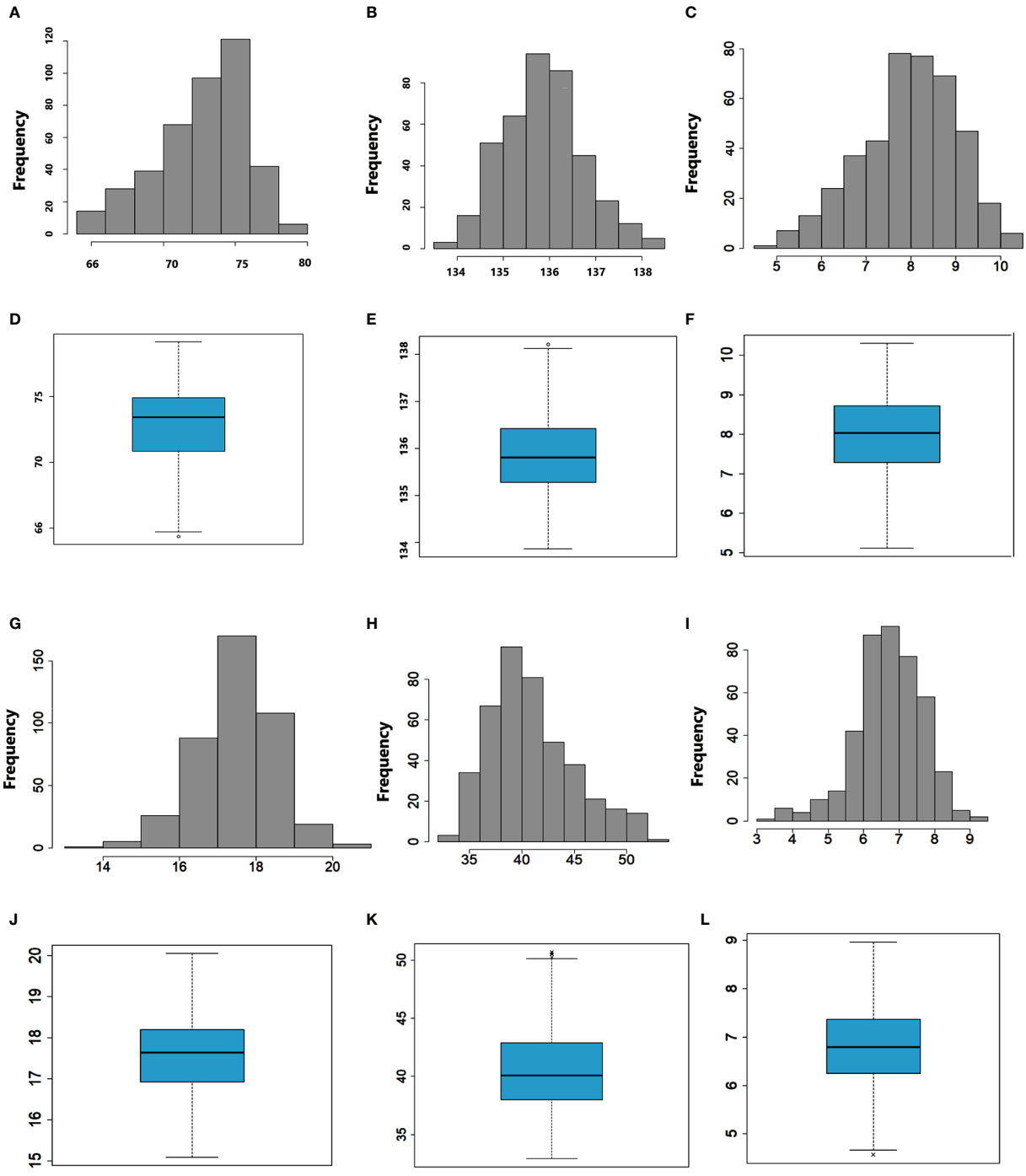


FIGURE 1
Frequency distribution and boxplots of DTH **(A, D)**, DTM **(B, E)**, SPL **(C, F)**, SPP **(G, J)**, TKW **(H, K)**, and GYD **(I, L)**.

Supplementary Figure 1). The fastest decrease of LD at cut-off (r^2 = 0.3) was observed on chromosome 7A. The r^2 values of marker pairs progressively declined as the physical distance between them increased on each chromosome (Supplementary Figure 1).

3.5 Principal component analysis, population structure, and kinship

The PCA scatter plot explained 92% (PC1 = 78.8% and PC2 = 13.2%) of the entire variation in the data set and grouped the

10.3389/fpls.2022.1009244

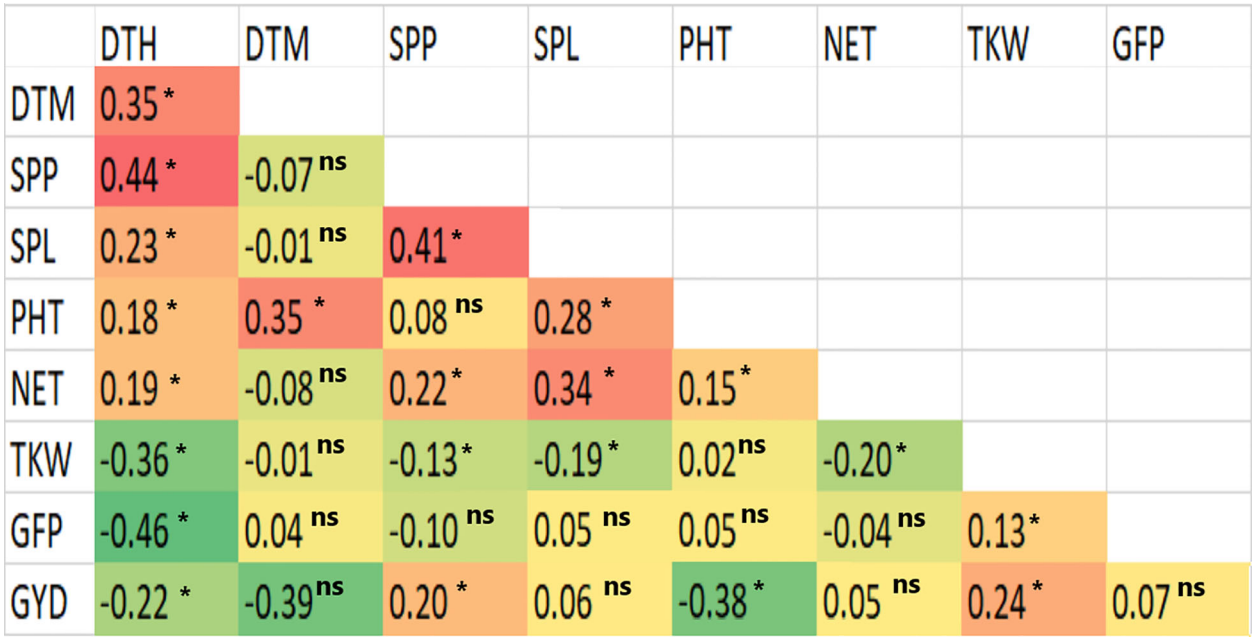


FIGURE 2
Computed correlation plots between pairs of phenotypic traits based on the best linear unbiased estimators of the nine traits measured in 420 durum wheat genotypes. N.B. *refers to significant at $p< 0.01$, and ns refers to non-significant at $p< 0.01$.

genotypes into two subpopulations (Figure 4A). Subpopulation 1 contained almost all modern cultivars, and subpopulation two included all landraces by showing clear grouping based on genetic background (Figures 4A, B).

A model-based Bayesian cluster analysis using STRUCTURE revealed that the optimal uppermost clear true ΔK value was obtained at best when K = 2, suggesting the 420 genotypes form two subpopulations (Figure 4D). Based on this clustering, cluster-1 comprised 348 landraces and one cultivar (85.5% of the genotype), and cluster-2 comprised 33 cultivars and 28 landraces (14.5% of the genotype). Analysis of admixture and purity using the STRUCTURE based on the Q value score (Q< 0.80 = admixture, and Q > 0.80 = pure genotypes) revealed that only 17 individuals (16 landraces and 1 cultivar) were classified as admixtures.

The STRUCTURE analysis revealed that 27 landraces were grouped with cultivars, and of these, three landraces (G242, G243, and G368) had a Q value of 1, which indicates 100% fitting the grouping with modern cultivars. Furthermore, one modern cultivar (G405), which was derived from related to landraces (according to their pedigree information), was grouped with landraces. The allelic

TABLE 2 The distribution of the 10,045 SNP markers across the entire durum wheat genome.

CHR[z]	NSPChr	GCR (bp)	SGRC (Mbp)	CHR	NSPChr	ROGC (bp)	SGRC (Mbp)
1A	757	1104472–585259074	584.2	1B	849	313555–681099620	680.8
2A	728	295475–774813964	774.5	2B	869	406084–789416853	789.4
3A	610	304055–746380464	746.1	3B	806	304239–746380464	746.1
4A	520	698412–736473645	735.8	4B	415	42526–674744571	674.7
5A	746	27537–667286510	667.3	5B	917	2555603–701346725	698.8
6A	643	591650–615260837	614.7	6B	753	2052283–698554772	696.5
7A	803	171878–727023089	726.9	7B	629	47368–721753586	721.7
A[a]	4807	na	4849.5	B[b]	5238	na	5008

[z] CHR, chromosome; [a]A, A genome; [b]B, B genome; NSPChr, Number of SNPs per chromosome; GCR, Genome coverage range; SGRC, Size of genomic region(s) covered; na, Not applicable.

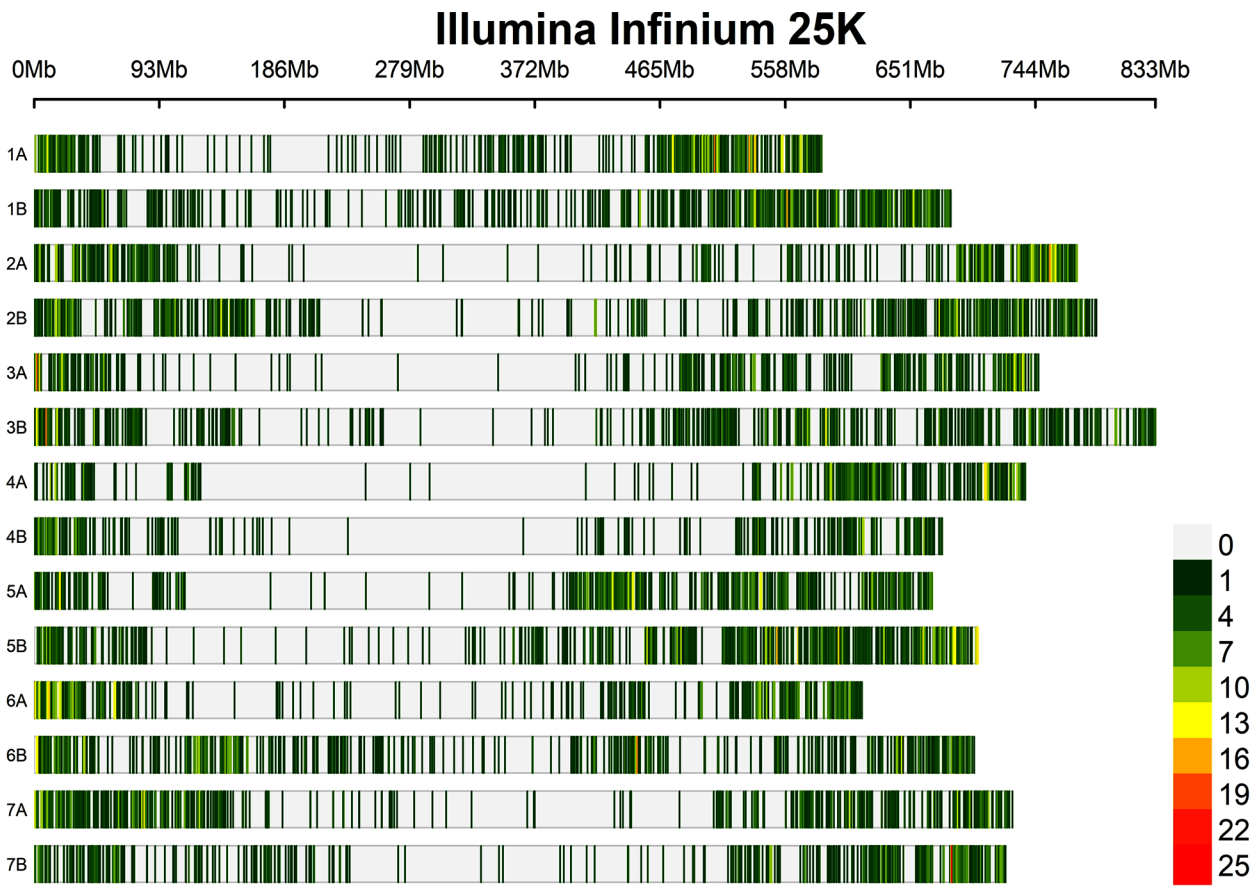


FIGURE 3
Distribution of the SNPs used in the present study on each durum wheat chromosome.

divergence between the two subpopulations inferred by STRUCTURE was 0.27. On average, the expected heterozygosity of subpopulation-1 (Cl - I) and subpopulation-2 (Cl - II) was 0.22 and 0.32, respectively. Subpopulation-1 had a mean F_{ST} value of 0.62, while subpopulation-2 had a mean F_{ST} value of 0.35, indicating high differentiation among the individuals of each population. Although a slight difference was observed, model-based Bayesian clustering and distance-based PCA similarly grouped the individuals into two subpopulations. The kinship matrix heatmap revealed familial relationships between the genotypes, which can be regarded as intermediate on average (Figure 4C).

3.6 GWAS scan of phenotypic traits

Considering all test locations and combined data over locations, GWAS was able to identify 179 significant MTAs for the nine traits (Supplementary Table 6). Of these, 23 MTAs were detected for DTH, 32 MTAs for DTM, 15 MTAs for GFP, 8 MTAs for GYD, 5 MTAs for NET, 19 MTAs for PHT, 26 MTAs for spike length (SPL), 12 MTAs for SPP and 39 MTAs for TKW. Using BLUEs of combined data across the five environments revealed 44 significant MTAs for the nine traits evaluated in this study. Further results and discussions (below) focus on these significant MTAs identified using the combined data across the five environments. The Manhattan and quantile-quantile (Q-Q) plots for each trait and environment are presented in Supplementary Figures 2A-E, respectively.

3.6.1 Marker trait association for phenological traits

For phenological traits (DTH, DTM, and GFP), 12 significant MTAs were identified from the GWAS of combined data from the five locations (Table 4). The GWAS scan for DTH detected six significant MTAs on chromosomes 1B, 2A, 5B, 6B, 7A, and 7B (Figure 5 and Table 4). The Q-Q plot showed that the data fitted the model well, and false positive MTAs were controlled. Among these MTAs, three were previously reported (Golabadi et al., 2011;

TABLE 3 A summary of linkage disequilibrium analysis for SNP marker pairs and the distribution of significant SNP pairs across each chromosome of each genome.

Chromosome	Total number of SNP pairs	Significant SNP marker pairs at $r^2 \geq 0.3$ ($p < 0.01$)	Average r^2	Average distance (Mbp [z])
1A	35,875	7,811 (21.8%)	0.20	20.4
1B	41,700	7,666 (18.4%)	0.16	21.1
2A	35,700	7,218 (20.2%)	0.17	28.3
2B	42,700	8,709 (20.4%)	0.17	23.8
3A	29,800	5,297 (17.8%)	0.16	32.9
3B	39,550	10,362 (26.2%)	0.21	27.4
4A	25,250	3,580 (14.2%)	0.14	38.4
4B	20,000	3,805 (19.0%)	0.17	44.9
5A	36,600	8,144 (22.3%)	0.19	23.7
5B	45,100	8,738 (19.4%)	0.17	20.1
6A	31,450	8,454 (26.9%)	0.22	25.6
6B	36,900	7,586 (20.6%)	0.18	24.6
7A	39,450	4,781 (12.1%)	0.12	23.9
7B	30,700	5,234 (17.0%)	0.15	30.8
A genome	234,125	45,284 (19.3%)	0.17	26.9
B genome	293,250	52,102 (17.8%)	0.17	25.8
Total	490,775	97,386 (19.8%)	0.17	26.8

[z] Mbp, Mega base pair.

Giraldo et al., 2016; Mangini et al., 2018), while three MTAs on the B genome (AX-109859693, wsnp_BE496986B_Ta_2_2, Ku_c24482_1132) were novel. The significant MTAs explained 1.1 to 22.2% of the total phenotypic variation in DTH. Among the significant MTAs for DTH, wsnp_BE496986B_Ta_2_2, AX-109859693, Ku_c24482_1132, and IACX11338 appeared significant in two or more test environments and hence can be regarded as stable MTAs.

The GWAS scan detected four significant MTAs across test environments for DTM. Of these MTAs, Kukri_rep_c73477_888 on chromosome 6A was previously reported (Mangini et al., 2018) and was detected in two environments. Three MTAs (Tdurum_contig49186_437 and Tdurum_contig12722_779 on chromosome 7A; and AX-109869840 on chromosome 6A) were likely to be potential new loci (Table 4 and Figure 5). The proportion of phenotypic variance explained by these four significant SNPs ranged from 2 to 33%.

GWAS revealed two significant MTAs for GFP on chromosomes 3B and 7B (Table 4 and Figure 5). The RAC875_c62223_86 MTA on 3B was previously reported (Giraldo et al., 2016). However, Kukri_c60966_261 on chromosome 7B was novel and detected repeatedly in two locations. RAC875_c62223_86 and Kukri_c60966_261 explained 4% and 31% of the variation in GFP obtained in the present study, respectively.

3.6.2 Marker trait association for plant architecture

The GWAS analysis revealed 14 MTAs significantly associated with plant architecture traits (Table 4 and Figure 6). For PHT, six significant MTAs were detected (Figure 6). Among these, four (AX-158602974, and AX-95259256 on chromosome 1B, wsnp_BE443745A_Ta_2 on chromosome 5A, and BS00091519_51 on chromosome 5B) were previously reported (Zhang et al., 2012; Mengistu et al., 2016; Roncallo et al., 2017). The other two MTAs (AX-95154560 and Tdurum_contig75127_589 on chromosomes 1B and 7B, respectively) were novel. The six significant MTAs explained 1.2 to 23% of the variation in PHT recorded in this study.

The MTA analysis for spike length (SPL) revealed five significant MTAs. Among the significant MTAs, Tdurum_contig45715_1246 on chromosome 1B was previously identified (Giraldo et al., 2016). The remaining four MTAs (Kukri_c17062_618 and Tdurum_contig76960_213 on chromosome 2A, Kukri_c3096_1411 on chromosome 2B, and AX-94615777 on chromosome 5A) are novel (Figure 6). The

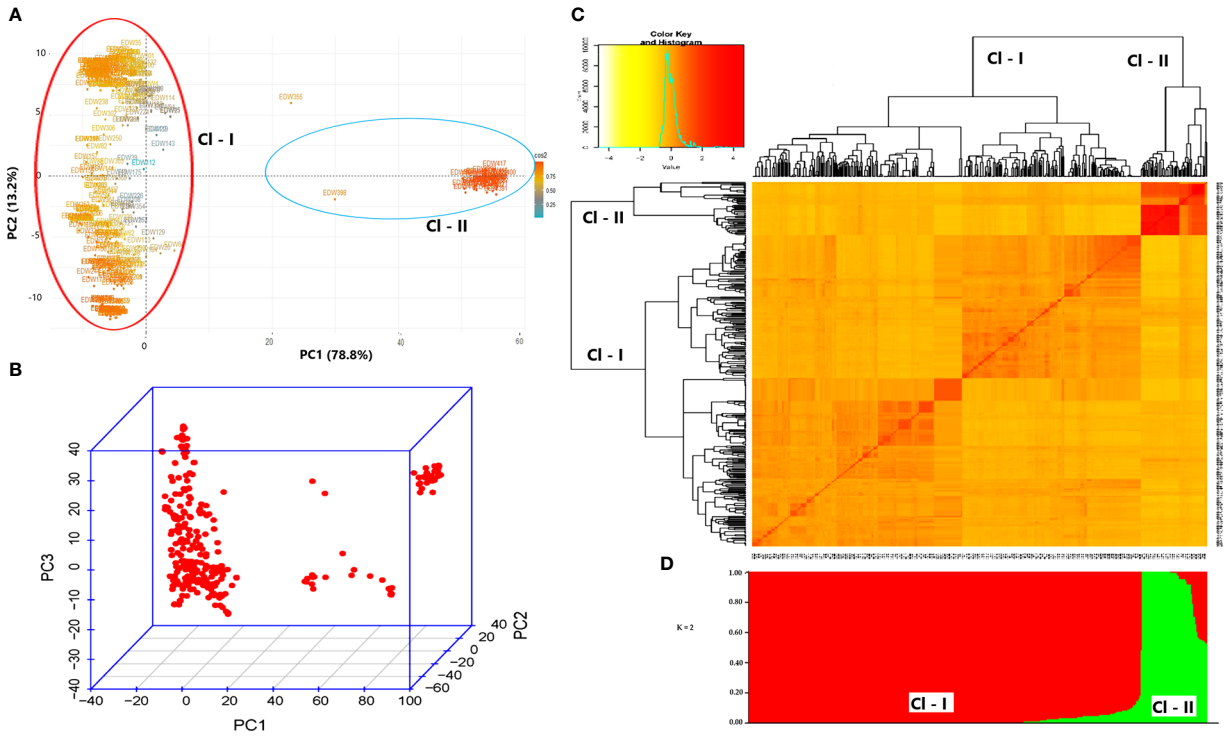


FIGURE 4
Principal component analysis **(A, B)**, kinship **(C)**, and population structure analysis **(D)**.

proportion of the variation in SPL elucidated by the significant MTAs varied from 2.75% to 12.3%. For NET, GWAS revealed two (GENE-0410_71 and AX-94782013) significant MTAs on chromosome 1B and 7B, respectively (Table 4 and Figure 6), which have not been previously reported. These MTAs explained 2.8% and 5.6% of the variation in NET, respectively.

3.6.3 Marker trait association for grain yield and related traits

GWAS for grain yield and yield-related traits evidenced 18 significant MTAs (Table 6 and Figure 7). The association scan for SPP resulted in seven significant MTAs on chromosomes 1B, 2B, 3A, 4A, and 7A. The proportion of phenotypic variance explained by the associated MTAs ranged from 1.1% to 17%. Of the seven significant MTAs for SPP, four (AX-89760660, Tdurum_contig25602_212, BS00110281_51, and AX-158591111) were previously reported (Golabadi et al., 2011; Mengistu et al., 2016; Soriano et al., 2016; Kidane et al., 2017a; Roncallo et al., 2017; Abu-Zaitoun et al., 2018; Mangini et al., 2018), whereas the remaining three (RAC875_c400_193, AX-158597411, and AX-94631122) SNPs are novel.

For GYD, GWAS revealed four significant MTAs on chromosomes 1B, 5A, 5B, and 7A. The proportion of phenotypic variance explained by the significant MTAs ranged from 1.74% (RAC875_c57656_170 on chromosome 7A) to 44.95% (IAAV3365 on chromosome 5A). Alleles of the high signal MTAs (locus IAAV3365, A/G alleles) had a highly significant effect on grain yield (Figure 8A). The genotypes carrying allele A had higher average grain yield across the five environments as compared to genotypes carrying allele G. RAC875_c57656_170 was previously reported for GYD (Maccaferri et al., 2014), whereas the remaining three MTAs (IAAV3365, RFL_Contig3481_1669 on chromosome 1B, and Excalibur_c51720_84 on chromosome 5B) were newly detected in the present study.

The genome-wide association analysis identified seven significant MTAs for TKW on chromosomes 1B, 3B, 5A, 6A, and 7A. The phenotypic variance explained by the associated SNPs ranged from 1.05% to 10.6%. Among the MTAs significantly associated with TKW, two (AX-158606713 and wsnp_Ex_rep_c66939_65371026 on chromosomes 1B and 7A, respectively) were previously reported. However, the other five MTAs (BS00071597_51, AX-158541767, RAC875_c41315_157, AX-158564275, and AX-94640059 on chromosomes 3B, 5A, 6A, and 7A, respectively) were novel. The effect of alleles on locus AX-158564275 (A/G alleles) revealed a highly significant difference in TKW (Figure 8B). The genotypes with the allele A had high TKW compared to genotypes carrying allele G.

10.3389/fpls.2022.1009244

TABLE 4 Summary of significant marker-trait associations for the nine traits revealed based on the combined data of the five locations on each durum wheat chromosome (CHR).

SNP (MTAs)	CHR	POS (bp)	P-value	MAF	Effect	PVE^Z	Trait
AX-158591262	7A	30075367	5.13281E-07	0.31	-0.51	1.62	DTH
AX-94884567	2A	756029643	5.54677E-08	0.16	0.77	1.1	
IACX11338	1B	522669860	1.43456E-06	0.09	-2.04	22.18	
AX-109859693	5B	5172617	3.78017E-09	0.05	1.16	5.13	
wsnp_BE496986B_Ta_2_2	6B	568039035	1.80668E-08	0.17	-0.68	5.28	
Ku_c24482_1132	7B	155935903	3.69626E-07	0.29	-0.66	1.67	
AX-109869840	6A	603461435	2.75361E-06	0.05	-0.74	21.07	DTM
Kukri_rep_c73477_888	2A	70572673	1.55589E-23	0.09	-2.54	32.88	
Tdurum_contig49186_437	7A	32558067	5.98292E-09	0.49	0.29	1.88	
Tdurum_contig12722_779	7A	44540113	4.95657E-07	0.07	0.41	4.45	
RAC875_c62223_86	3B	763192473	1.31126E-07	0.32	-0.53	3.58	GFP
Kukri_c60966_261	6B	693337622	5.50108E-07	0.08	-1.41	30.67	
RFL_Contig3481_1669	1B	4043159	5.87000E-06	0.10	0.02	5.45	GYD
Excalibur_c51720_84	7A	709197555	6.71000E-06	0.05	0.02	15.75	
RAC875_c57656_170	7A	614197852	1.75039E-06	0.46	0.17	1.74	
IAAV3365	5A	548344620	1.87701E-10	0.06	-0.79	44.95	
GENE-0410_71	1B	523053033	1.78840E-07	0.44	0.22	5.61	NET
AX-94782013	7B	604310198	3.10157E-06	0.09	-0.25	2.79	
AX-158602974	1B	580658793	5.16606E-11	0.06	-4.33	2.68	PHT
BS00091519_51	5B	5174649	2.79112E-06	0.05	-2.03	9.64	
AX-158521163	1B	669849432	3.25148E-06	0.06	-2.10	1.2	
AX-95259256	1B	629504430	2.07679E-07	0.12	-1.52	4.51	
wsnp_BE443745A_Ta_2_1	5A	439542987	7.51027E-08	0.37	1.12	2.59	
AX-95154560	6B	120830636	5.51278E-08	0.05	2.73	23.76	
Tdurum_contig75127_589	7B	697951769	5.0922E-08	0.06	-3.88	8.34	
Tdurum_contig45715_1246	1B	314321199	6.74631E-07	0.48	-0.26	3.62	SPL
Kukri_c17062_618	2A	522595273	4.96382E-07	0.28	0.17	2.75	
Tdurum_contig76960_213	2A	492195805	3.17135E-07	0.13	-0.63	12.3	
Kukri_c3096_1411	2B	314134332	3.40111E-07	0.22	0.28	3.02	
AX-94615777	5A	529858969	2.51717E-07	0.45	0.15	1.01	
Tdurum_contig25602_212	2B	546442999	3.25386E-07	0.24	0.22	1.76	SPP
AX-158591111	7A	33518205	7.03418E-08	0.24	-0.24	1.85	
BS00110281_51	4A	724872914	1.51333E-06	0.06	0.41	4.5	
AX-89760660	1B	519060573	8.65163E-09	0.07	-0.49	16.82	
RAC875_c400_193	1B	1547605	8.21951E-07	0.10	-0.25	1.53	
AX-158597411	2B	99223728	2.38152E-06	0.49	0.24	2.27	
AX-94631122	3A	723577013	1.23025E-09	0.23	0.31	1.12	

(Continued)

TABLE 4 Continued

SNP (MTAs)	CHR	POS (bp)	P-value	MAF	Effect	PVE^{Z}	Trait
AX-158606713	1B	546979073	1.26545E-06	0.26	-0.67	1.05	
wsnp_Ex_rep_c66939_65371026	7A	6480158	1.99623E-06	0.18	-0.81	1.72	
BS00071597_51	3B	803879943	7.43197E-09	0.20	1.44	7.93	
AX-158541767	3B	61267921	2.83169E-07	0.07	-1.82	10.6	TKW
RAC875_c41315_157	5A	431829169	3.78422E-06	0.10	-1.28	5.65	
AX-158564275	6A	528989018	3.23204E-10	0.09	-1.57	5.45	
AX-94640059	7A	686968079	4.86263E-06	0.44	0.65	1.23	

Chr, Chromosome; POS, Physical position of SNP; bp, Base pair; MAF, Minor allele frequency; PVE, Phenotypic variance explained; DTM, Days to heading; DTM, Days to maturity; GFP, Grain filling period; NET, Number of effect tillers per plant; SPL, Spike length; PHT, Plant height; SPP, Number of spikelets per spike; TKW, Thousand kernel weight; GYD, Grain yield.

3.6.4 Identification of putative novel MTAs and their underlying candidate genes

According to the LD decay information for each chromosome, a genomic region of ten Mbp around each significant SNP (five Mbp downstream and five Mbp upstream of the significant SNP) is considered to be a QTL. Significant SNPs within the ranges of 10 Mbp apart are considered to refer to the same QTL. Based on this approach, 37 QTLs were identified for the 44 significant MTAs (Table 5). The names of these QTLs (*q.gwas.01* to *q.gwas.*37) are provided in the first column of Table 5. Among the 37 QTLs, 16 were located in or near genomic regions previously reported for the corresponding traits (Golabadi et al., 2011; Maccaferri et al., 2014; Giraldo et al., 2016; Mengistu et al., 2016; Soriano et al., 2016; Kidane et al., 2017a; Abu-Zaitoun et al., 2018; Mangini et al., 2018; Roncallo et al., 2018), while 21 were novel (Table 5). Two QTLs for SPP (*q.gwas.15* and *q.gwas.20*) and one QTL (*q.gwas.08*) for TKW were previously described based on Ethiopian durum wheat germplasm (Mengistu et al., 2016; Kidane et al., 2017a). Genomic regions for five putative QTLs (*q.gwas.01, q.gwas.02, q.gwas.22, q.gwas.24*, and *q.gwas.30*) overlap with more than one trait evaluated in this study (Table 5). For example, four significant MTAs for DTH, DTM, SPP, and TKW were co-localized and hence were considered to be referring to the same QTL (*q.gwas.30*) (Table 5). The analysis of the sequences of these putative QTLs genomic regions based on durum wheat, the reference genome at the Ensemble Plants database, led to the identification of 774 potential candidate genes (Supplementary Table 7).

The significance of the candidate genes was evaluated by reviewing previously published genomic regions associated with the traits targeted in the present study (Zhang et al., 2012; Maccaferri et al., 2016; Maccaferri et al., 2019; Kidane et al., 2019; Mazzucotelli et al., 2020; Zhao et al., 2020). This resulted in 32 genes related to eight of the nine target traits in durum wheat (Table 6). The putative candidate genes *TRITD7Av1G01175*, *TRITD7Av1G017240*, and *TRITD7Av1G017550* (all located on chromosome 7A), which encode Growth-regulating factor, Zinc-finger CCCH domain protein TE, and NAC domain-containing protein, respectively, were associated with DTH and DTM. The genes *TRITD1Bv1G168480 and TRITD7Bv1G057630* encode WRKY transcription factor and Flowering Locus T/Terminal Flower 1-like protein, respectively, were reported to regulate DTH. The *TRITD1Bv1G000110* gene, on chromosome 1B, which encodes Tryptophan aminotransferase-related protein 2, was shown to be associated with SPP and GYD (Table 6).

4 Discussion

This study used GWAS to define durum wheat genomic regions associated with phenological, plant architecture, grain yield, and yield-related traits. Furthermore, analyses of population structure and linkage disequilibrium were carried out to increase the efficiency of detecting reliable marker-trait associations as well as identifying the genetic basis of those associations. The present study utilized a large number of diverse durum wheat landraces and cultivars, which were grown across diverse environments in Ethiopia. This facilitated the identification of novel SNP loci associated with nine durum wheat phenotypic traits, including grain yield and grain yield-related traits. The present study findings have significant implications for both the development of molecular markers for genomics-led breeding and for providing new insights into the architecture of genomic regions regulating various traits of interest in durum wheat. These could facilitate the improvement of grain yield and other desirable characteristics to support global food security.

To meet the growing demand for durum wheat grains as well as the challenges to its production brought about by the expanding environmental changes, it is imperative that the genetic resources of durum wheat, including landraces, modern cultivars, and breeding lines, be effectively utilized for breeding new cultivars (Kankwatsa et al., 2017; Maccaferri et al., 2019b; Kumar et al., 2020; Mazzucotelli et al., 2020). As such, identifying loci that regulate desirable traits in breeding

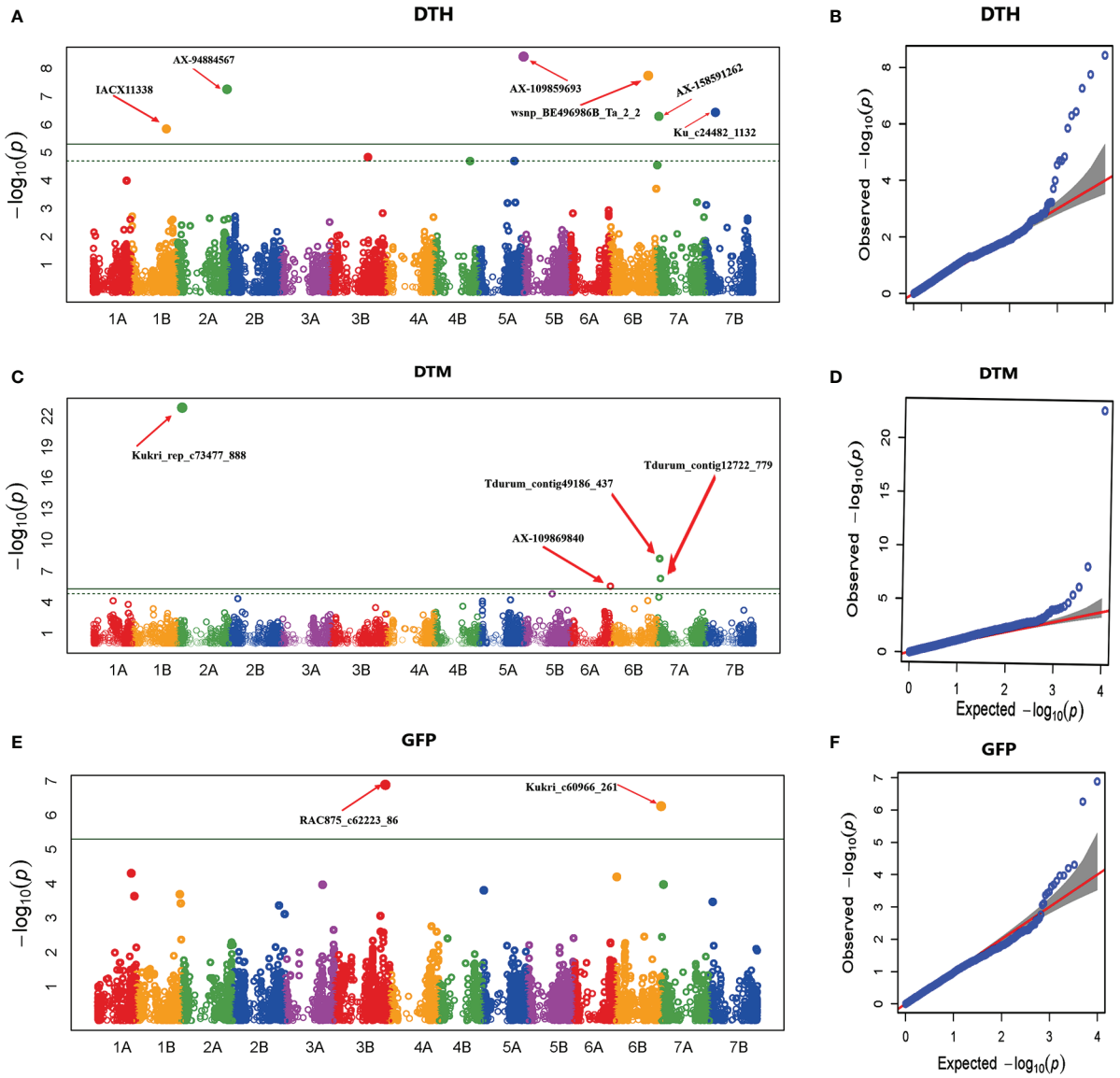


FIGURE 5
Manhattan and Q-Q plots of GWAS scan for phenological traits generated based on combined data from five locations. DTH **(A, B)**, DTM **(C, D)**, and GFP **(E, F)**. For the Manhattan plots, the y-axis represents -log10 (p) of the traits, while the x-axis represents the relative positions of the SNP markers on each chromosome. DTH, Days to heading; DTM, Days to maturity; GFP, Grain-filling period.

programs helps to develop markers for marker-assisted breeding, thus contributing to food security (Garcia et al., 2019; Wang et al., 2019; Mérida-García et al., 2020).

The present study revealed highly significant contributions of genotypes, environments, and genotype by environment interactions to the phenotypic variations of the target traits ($p< 0.001$), which is consistent with the results of previous research on durum wheat (Mengistu et al., 2015; Mohammadi et al., 2018; Mekonnen et al., 2021). The observed high genotypic variance, genotypic coefficient of variation, and broad-sense heritability for TKW and GFP, strongly suggest that their variation is mainly due to heritable genetic differences among the landraces and cultivars. There was a low genotypic variance and genotypic coefficient of variation for GYD, indicating the challenges associated with improving this trait. Nevertheless, moderate to a high level of broad-sense heritability were recorded for all traits, meaning that a significant part of the observed variation is heritable and that the results agree with previous findings in durum wheat (Sukumaran et al., 2018; Alemu et al., 2020a).

The present study found that GYD had a moderate but significant ($p< 0.01$) positive correlation with SPP, and TKW, indicating that the simultaneous selection of desirable characteristics of these traits could lead to the improvement of

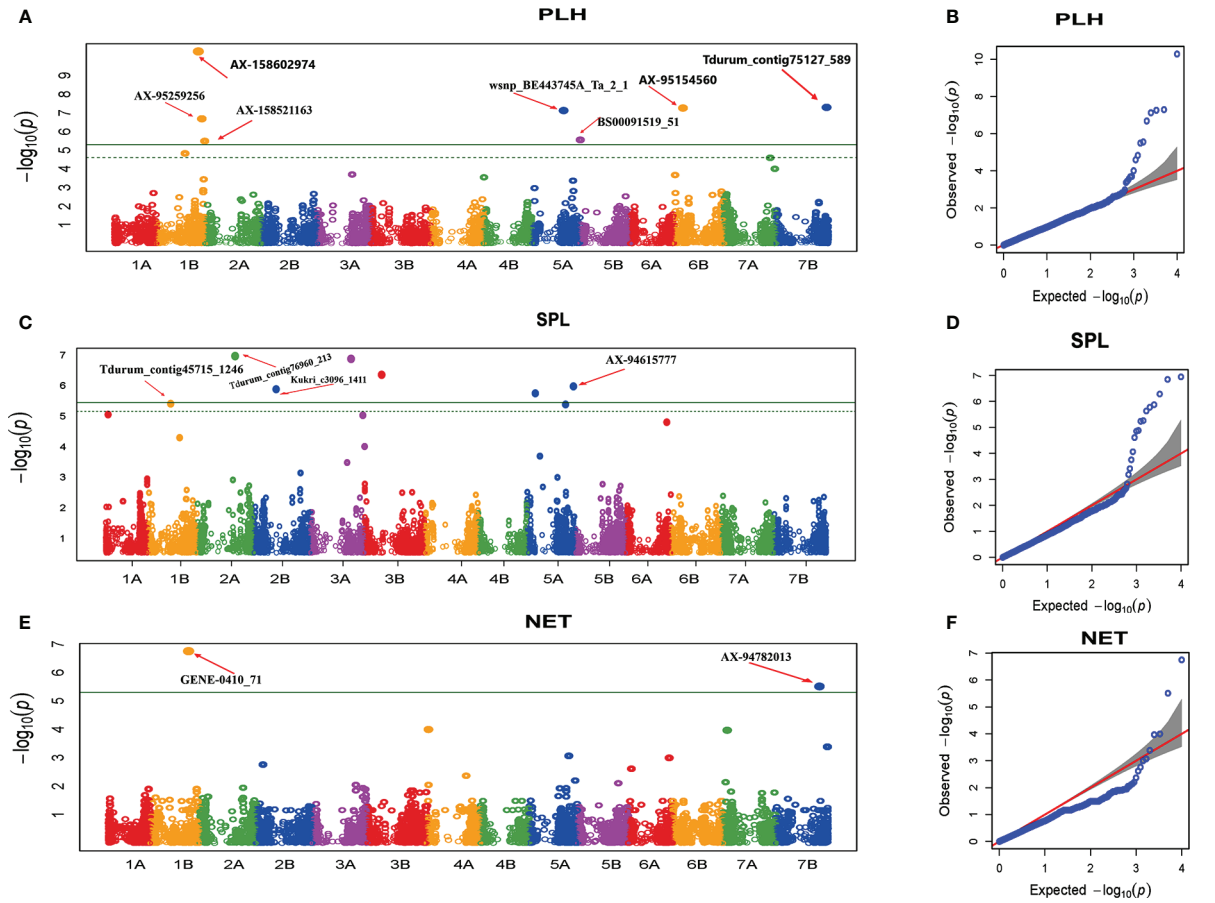


FIGURE 6
Manhattan and Q-Q plots of GWAS scan for plant architecture traits generated based on the combined data from five locations. PHT **(A, B)**, SPL **(C, D)**, NET **(E, F)**. For the Manhattan plots, the y-axis represents -log10 (p) of the traits, while the x-axis represents the relative positions of the SNP markers on each chromosome. PHT, Plant height; SPL, Spike length; NET, Number of effective tillers per plant.

grain yield in this crop. However, GYD negatively correlated with DTH and PHT, indicating that late-heading genotypes generally have lower grain yield than early-heading types. However, the early-heading types appear to have a more extended grain-filling period, as a very low but significant positive correlation was obtained for GYD versus DTM. TKW exhibited a moderate positive correlation with GYD, and GFP, implying that direct improvement of these traits may improve the former, which contributes to enhancing GYD. Conversely, TKW had a negative relation with DTH and DTM. Thus late-maturing cultivars will have a relatively low TKW.

The LD among the SNP marker pairs showed a sharp decline within the physical distance ranging from 2.02 to 10.4 Mbp, with an average of 4.26 Mbp. This decline in LD is far below the results of previous research using Ethiopian durum wheat landraces (Alemu et al., 2021b), which reported an average physical distance of 69.1 Mbp. Similarly, Mekonnen et al., (2021) found a higher mean LD decay (31.44 Mbp, r^2 = 0.2) in their study on diverse Ethiopian bread wheat germplasm. This disparity could arise due to the type and density of markers, genomic regions the markers cover, and differences in the sample used in these studies. However, Fayaz et al., 2019 found a low LD decay (2–3 cM) of the A and B sub-genomes using Iranian durum wheat landraces at a critical r^2 = 0.11. Likewise, Rufo et al., (2019) noted an LD decay ranging from 1 to 9 cM on A and B genomes from landraces and released cultivars of Mediterranean wheat. The fastest LD decay rate of an average physical distance of 2.02 Mbp was recorded for chromosome 7A.In contrast, the slowest was recorded for chromosome 4A (10.04 Mbp), which indicates the differences in recombination rates among different genomic regions of different chromosomes. On average, the A genome showed a more rapid LD decay than the B genome (Supplementary Table 4), and more substantial selection pressure could be partly caused in the A genome than in the B genome (Liu et al., 2019; Kumar et al., 2020). This result most likely confirms the impact of

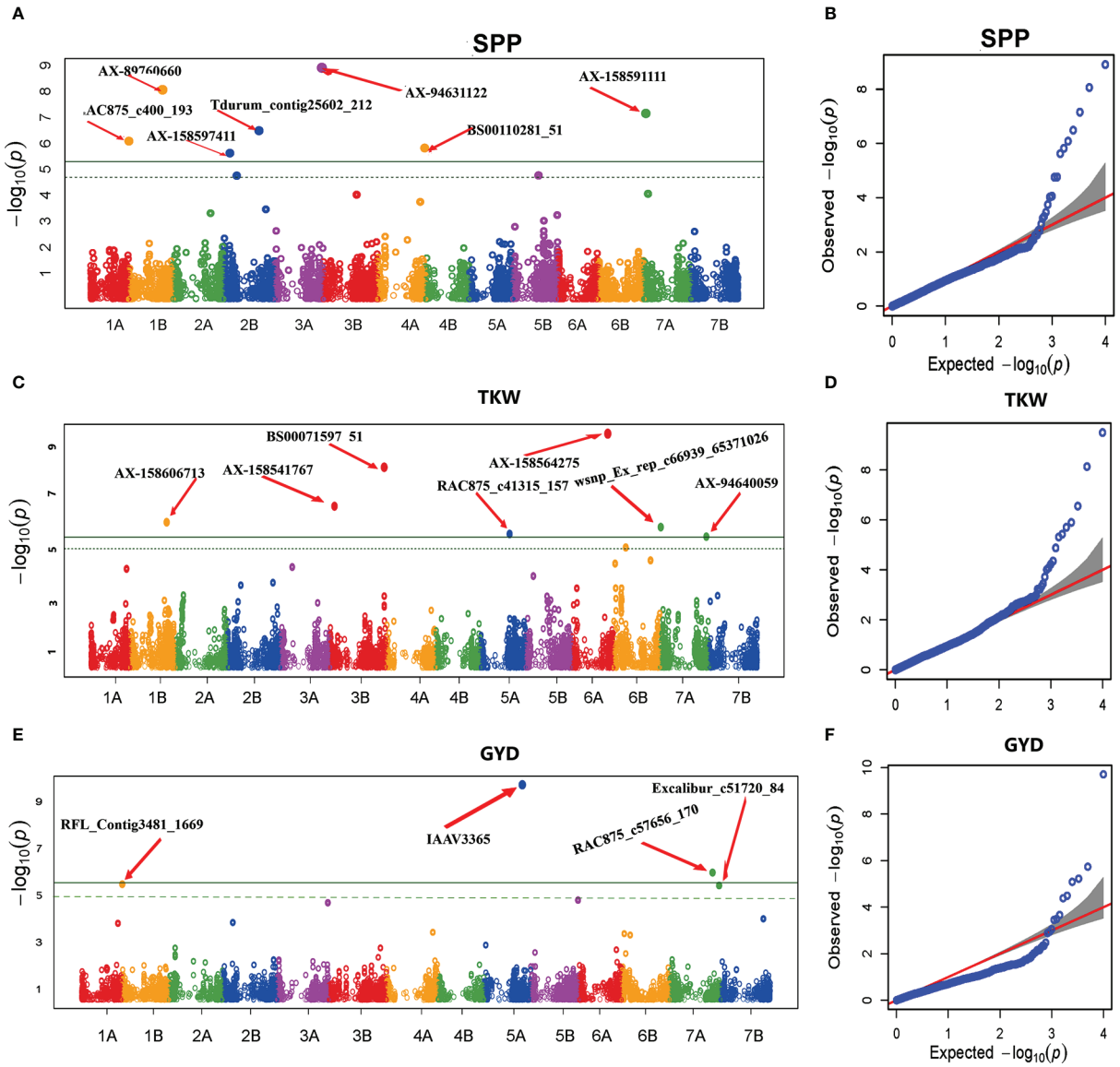


FIGURE 7
Manhattan and Q-Q plots of GWAS scan for SPP **(A, B)**, TKW **(C, D)**, and GYD **(E, F)** generated based on the combined data from five locations. For the Manhattan plots, the y-axis represents -log10 (p) of the traits, while the x-axis represents the relative positions of the SNP markers on each chromosome. SPP, Number of spikelets per spike; TKW, Thousand kernel weight; GYD, Grain yield.

genetic drift, mutation, gene flow, recombination, the pressure of population selection, and historical events on both A and B genomes (Fayaz et al., 2019).

The population clustering inferred by STRUCTURE and PCA divided the genotypes into two sub-populations, similar to the results of earlier research (Wang et al., 2019; Alemu et al., 2020b; Kumar et al., 2020; Mekonnen et al., 2021). Based on Q-score values of STRUCTURE analysis (Q > 0.80), 96% of the landraces were pure, and 4% of the genotype were admixtures. The kinship matrix was used to estimate the family relatedness and to confirm the relation within the genotypes. Hence, the cumulative results from STRUCTURE, PCA, and kinship suggest adjusting the GWAS model to avoid bias arising from spurious associations, thereby reducing false-positive associations arising from co-ancestry. Moreover, FarmCPU, a robust statistical model for GWAS, adequately accounted for the spurious associations that arose from population structure, cryptic relatedness, and marker effects, as shown by Q-Q plots. Based on the five-test sites´ mean data, the GWAS revealed 44 MTAs. The SNPs associated with the target traits were distributed across the whole chromosome except chromosome 1A, which did not bear any significant MTAs.

Using GWAS, different genomic regions associated with grain yield were identified in the present study. The putative

10.3389/fpls.2022.1009244

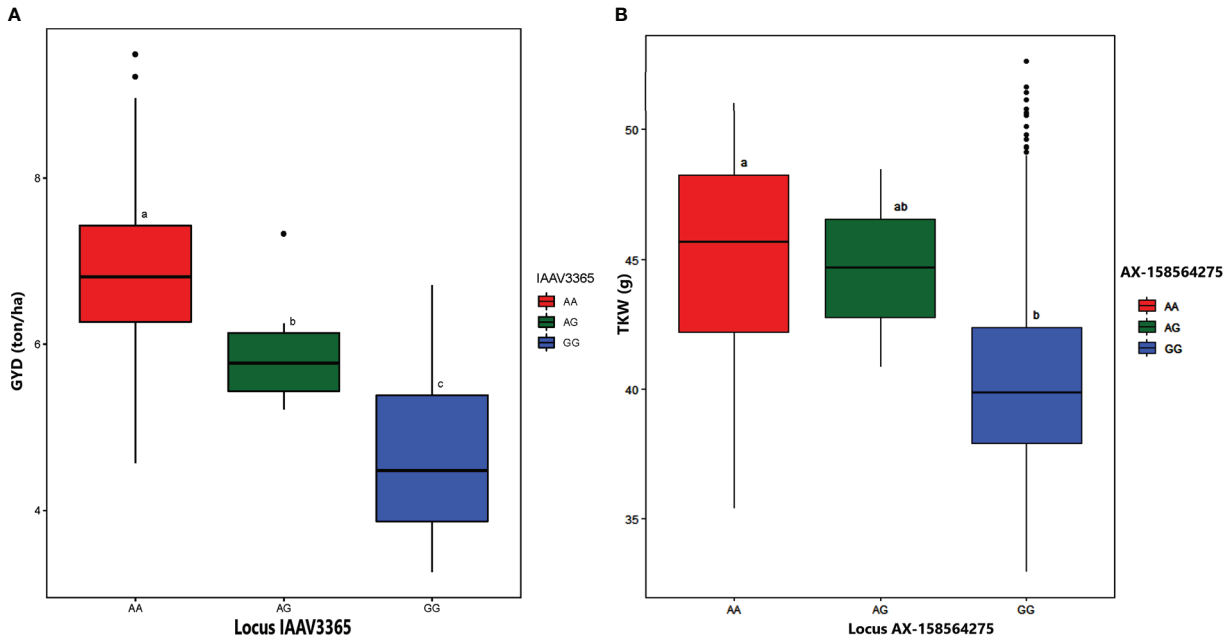


FIGURE 8
Boxplot depicting the effect of alleles located on locus IAAV3365 on grain yield **(A)** and locus AX-158564275 on thousand kernel weight **(B)**. The estimated mean BLUEs value was used to generate the boxplot to reveal the effects of respective alleles on grain yield and thousand kernel weight. Tukey´s HSD (Honestly Significant Difference) test was applied to see the differences of alleles in 420 durum wheat genotypes. The letters above the boxplot indicate the significant differences among the mean performance of genotypes carrying each allele at a probability level of $p< 0.05$.

TABLE 5 Summary of putative quantitative trait loci (QTLs) identified for the nine phenotypic traits analyzed in the present study using Ethiopian durum wheat germplasm.

Putative QTL	Associated SNP	Chr	SNP position (bp)	TAPQTL
q.gwas.01[a]	IACX11338	1B	522669860	DTH
q.gwas.02 [a]	RFL_Contig3481_1669	1B	4043159	GYD
q.gwas.01 [a]	GENE-0410_71	1B	523053033	NET
q.gwas.03 [a]	AX-158602974	1B	580658793	PHT
q.gwas.04 [a]	AX-158521163	1B	669849432	PHT
q.gwas.05	AX-95259256	1B	629504430	PHT
q.gwas.06 [a]	Tdurum_contig45715_1246	1B	314321199	SPL
q.gwas.02	RAC875_c400_193	1B	1547605	SPP
q.gwas.07	AX-89760660	1B	519060573	SPP
q.gwas.08 [a]	AX-158606713	1B	546979073	TKW
q.gwas.09 [a]	AX-94884567	2A	756029643	DTH
q.gwas.10	Kukri_rep_c73477_888	2A	70572673	DTM
q.gwas.11	Tdurum_contig76960_213	2A	492195805	SPL
q.gwas.12	Kukri_c17062_618	2A	522595273	SPL
q.gwas.13	Kukri_c3096_1411	2B	314134332	SPL

(Continued)

10.3389/fpls.2022.1009244

TABLE 5 Continued

Putative QTL	Associated SNP	Chr	SNP position (bp)	TAPQTL
q.gwas.14 [a]	AX-158597411	2B	99223728	SPP
q.gwas.15	Tdurum_contig25602_212	2B	546442999	SPP
q.gwas.16	AX-94631122	3A	723577013	SPP
q.gwas.17 [a]	RAC875_c62223_86	3B	763192473	GFP
q.gwas.18	AX-158541767	3B	61267921	TKW
q.gwas.19	BS00071597_51	3B	803879943	TKW
q.gwas.20 [a]	BS00110281_51	4A	724872914	SPP
q.gwas.21	IAAV3365	5A	548344620	GYD
q.gwas.22	wsnp_BE443745A_Ta_2_1	5A	439542987	PHT
q.gwas.23	AX-94615777	5A	529858969	SPL
q.gwas.22	RAC875_c41315_157	5A	431829169	TKW
q.gwas.24 [a]	AX-109859693	5B	5172617	DTH
q.gwas.24 [a]	BS00091519_51	5B	5174649	PHT
q.gwas.25 [a]	AX-109869840	6A	603461435	DTM
q.gwas.26	AX-158564275	6A	528989018	TKW
q.gwas.27	wsnp_BE496986B_Ta_2_2	6B	568039035	DTH
q.gwas.28	Kukri_c60966_261	6B	693337622	GFP
q.gwas.29	AX-95154560	6B	120830636	PHT
q.gwas.30 [a]	AX-158591262	7A	30075367	DTH
q.gwas.30 [a]	Tdurum_contig49186_437	7A	32558067	DTM
q.gwas.31	Tdurum_contig12722_779	7A	44540113	DTM
q.gwas.32	RAC875_c57656_170	7A	614197852	GYD
q.gwas.33 [a]	wsnp_Ex_c16045_24471413	5B	685974689	GYD
q.gwas.30 [a]	AX-158591111	7A	33518205	SPP
q.gwas.30 [a]	wsnp_Ex_rep_c66939_65371026	7A	6480158	TKW
q.gwas.34	AX-94640059	7A	686968079	TKW
q.gwas.35	Ku_c24482_1132	7B	155935903	DTH
q.gwas.36	AX-94782013	7B	604310198	NET
q.gwas.37	Tdurum_contig75127_589	7B	697951769	PHT

[a]Previously identified QTLs, Chr, chromosome; DTH, Days to heading; DTM, Days to physiological maturity; GFP, Grain filling; NET, Number of effective tillers per plant; GFP, Grain filling period; PHT, Plant height; SPL, Spike length; SPP, Number of spikelets per spike; TKW, Thousand kernel weight; GYD, Grain yield; TAPQTL, Traits associated with Putative QTL.

QTLs identified for this trait are *q.gwas.02* (0.3 – 5.8 Mbp) on chromosome 1B, *q.gwas.21* (544.7 – 554.9 Mbp) on chromosome 5A, *q.gwas.32* (609.2 – 619.2 Mbp) on chromosome 7A, and *q.gwas.33* (703.3 – 714.2 Mbp) on chromosome 7A. Among them, *q.gwas.02*, *q.gwas.21*, and *q.gwas.32* are novel QTLs, as these genomic regions have not previously been reported for their association with grain yield. The putative QTL *q.gwas.33* on chromosome 5B is co-localized within the same genomic region of a QTL reported by Maccaferri et al., (2014); Maccaferri et al., (2016) for grain yield and total root numbers, respectively using durum wheat recombinant inbred lines. The QTL regions of *q.gwas.33* is also identified for spikes per plant (Mengistu et al., 2016), kernel Fe content (Velu et al., 2017), kernels per spikelets (Peng et al., 2003), fusarium head blight resistance (Ghavami et al., 2011), yellow rust resistance (Liu et al., 2017a), and stem rust resistance

10.3389/fpls.2022.1009244

TABLE 6 Summary of selected genes associated with some of the putative QTLs identified in the present study.

S/N	SNP	Chr	PQAG	Gene ID	GSS	GSE	Description of gene	Trait
1	IACX11338	1B	*q.gwas.01*	*TRITD1Bv1G168480*	523422292	523424581	WRKY transcription factor	
2	IACX11338	1B	*q.gwas.01*	*TRITD1Bv1G169970*	526873942	526876463	WRKY transcription factor	
3	IACX11338	1B	*q.gwas.01*	*TRITD1Bv1G170060*	527017659	527018012	WRKY DNA-binding protein 39 G	
4	AX-158591262	7A	*q.gwas.30*	*TRITD7Av1G011750*	20905388	20908162	Growth-regulating factor	
5	AX-158591262	7A	*q.gwas.30*	*TRITD7Av1G017240*	30780092	30782694	Zinc finger CCCH domain protein TE?	DTH
6	AX-158591262	7A	*q.gwas.30*	*TRITD7Av1G017550*	31546753	31548244	NAC domain-containing protein, putative	
7	Ku_c24482_1132	7B	*q.gwas.35*	*TRITD7Bv1G056500*	157854628	157855010	Seed maturation protein LEA 4	
8	Ku_c24482_1132	7B	*q.gwas.35*	*TRITD7Bv1G056720*	158562417	158562866	Zinc finger family protein	
9	Ku_c24482_1132	7B	*q.gwas.35*	*TRITD7Bv1G057630*	162519251	162520668	FLOWERING LOCUS T/TERMINAL FLOWER 1-like protein	
10	Ku_c24482_1132	7B	*q.gwas.35*	*TRITD7Bv1G058480*	164897716	164902189	Phosphate transporter PHO1-like protein	
11	AX-109869840	6A	*q.gwas.25*	*TRITD6Av1G220960*	603258174	603260717	Ethylene receptor	
12	Tdurum_contig49186_437	7A	*q.gwas.30*	*TRITD7Av1G011750*	20905388	20908162	Growth-regulating factor	DTM
13	Tdurum_contig49186_437	7A	*q.gwas.30*	*TRITD7Av1G017240*	30780092	30782694	Zinc finger CCCH domain protein TE?	
14	Tdurum_contig49186_437	7A	*q.gwas.30*	*TRITD7Av1G017550*	31546753	31548244	NAC domain-containing protein, putative	
15	Kukri_c60966_261	6B	*q.gwas.28*	*TRITD6Bv1G226900*	693249887	693252855	Receptor protein kinase, Putative	GFP
16	AX-158602974	1B	*q.gwas.03*	*TRITD1Bv1G189370*	580660944	580663298	Calcineurin B-like protein	
17	AX-158602974	1B	*q.gwas.03*	*TRITD1Bv1G189570*	580988293	580988886	Receptor-like protein kinase	PHT
18	AX-158602974	1B	*q.gwas.03*	*TRITD1Bv1G191400*	585136543	585142125	Zinc finger protein	
19	BS00091519_51	5B	*q.gwas.24*	*TRITD5Bv1G001780*	5178666	5198798	Cytochrome P450-like protein	
20	Kukri_c17062_618	2A	*q.gwas.12*	*TRITD2Av1G189490*	526757053	526762842	Acyl-CoA N-acyltransferase	
21	Kukri_c17062_618	2A	*q.gwas.12*	*TRITD2Av1G190600*	529454657	529455487	Ring finger protein, Putative	SPL
22	Kukri_c3096_1411	2B	*q.gwas.13*	*TRITD2Bv1G109560*	316041878	316042468	E3 ubiquitin-protein ligase	
23	RAC875_c400_193	1B	*q.gwas.02*	*TRITD1Bv1G000110*	327500	328114	Tryptophan aminotransferase-related protein 2	
24	AX-89760660	1B	*q.gwas.07*	*TRITD1Bv1G166820*	518699664	518701642	Zinc finger CCCH domain-containing protein 4	
25	AX-89760660	1B	*q.gwas.07*	*TRITD1Bv1G167110*	519057378	519065351	UDP-GLUCOSE PYROPHOSPHORYLASE 1	SPP
26	Tdurum_contig25602_212	2B	*q.gwas.15*	*TRITD2Bv1G184650*	545766407	545768972	ethylene-responsive transcription factor	
27	AX-94631122	3A	*q.gwas.16*	*TRITD3Av1G275580*	723577014	723578195	E3 ubiquitin-protein ligase	
28	AX-158606713	1B	*q.gwas.08*	*TRITD1Bv1G176830*	544878459	544879181	Ethylene-responsive factor-like transcription factor	
29	AX-158606713	1B	*q.gwas.08*	*TRITD1Bv1G177060*	545904571	545908993	E3 ubiquitin-protein ligase	TKW
30	AX-158606713	1B	*q.gwas.08*	*TRITD1Bv1G177540*	547614469	547614996	Blue copper protein	
31	RFL_Contig3481_1669	1B	*q.gwas.02*	*TRITD1Bv1G000110*	327500	328114	Tryptophan aminotransferase-related protein 2	GYD
32	IAAV3365	5A	*q.gwas.21*	*TRITD5Av1G205000*	550462923	550477847	ABC transporter	

Chr, chromosome; GSS, Gene Sequence starts; GSE, Gene sequence ends; APQ, Associated putative QTL; DTH, Days to heading; DTM, Days to physiological maturity; GFP, Grain filling period; PHT, Plant height; SPL, Spike length; SPP, Number of spikelets per spike; TKW, Thousand kernel weight; GYD, Grain yield.

(Letta et al., 2014). This QTL *q.gwas.33* is also associated with genes *TRITD5BvG245710* (myb-like protein X), *TRITD5Bv1G246830* (KH domain containing protein), *TRITD5Bv1G247760* (NBS-LRR disease resistance protein –like protein) and *TRITD5Bv1G246270* (Glycosyltransferase).

The QTL region of *q.gwas.02* (RFL_Contig3481_1669) for GYD, identified in this study, overlaps with QTLs for several traits such as total root number and length (Maccaferri et al., 2016), grain protein content and concentration (Suprayogi et al., 2009), spikes per plant (Mengistu et al., 2016), heading date (Maccaferri et al., 2008), grain filling period (Soriano et al., 2016), semolina yellowness (Colasuonno et al., 2017), grain yield per spike and grain yield (Roncallo et al., 2018), grain protein content (Giraldo et al., 2016) and fusarium head blight resistance (Ghavami et al., 2011). The overlapping of QTLs for several important traits in this genomic region indicates its significance in future durum wheat breeding for grain yield and end-use quality traits. The genomic region corresponding to QTL *q.gwas.21* on chromosome 5A (MTA for IAAV3365 SNP and GYD) is a novel major QTL for grain yield, explaining the largest proportion of phenotypic variance (r^2 = 44.95%) as compared to all other putative QTLs reported here. This is a highly significant result of this study, which needs to be validated through further research, including fine mapping to pinpoint the gene(s) responsible for this QTL. Interestingly, the genomic region of this QTL overlaps with previously identified QTL for number of kernels per spike (Kidane et al., 2017a), yellow rust resistance (Liu et al., 2017b), threshing time (Tzarfati et al., 2014), leaf rust resistances (Aoun et al., 2016), and total root number (Maccaferri et al., 2016). Therefore, this genomic region is a key target region for the improving of durum wheat, for grain yield and threats of wheat arising due to the impacts of climate change. The *TRITD5Av1G205000* (an ABC transporter) gene is one of the potential candidate genes behind this QTL (*q.gwas.21*). This is because previous research indicated that ABC transporter genes affect grain formation in wheat during heading and also modulate the ripening of the heads (Wanke and Üner Kolukisaoglu, 2010; Walter et al., 2015).

The present study identified several novel QTLs for grain yield-related traits, SPP and TKW. Additional MTAs that confirmed previously identified genomic regions were also detected for these traits. The three novel putative QTLs for SPP are *q.gwas.07* (721.4–725.8 Mbp) on chromosome 1B, *q.gwas.15* (541.4–551.4 Mbp) on chromosome 2B and *q.gwas.16* (721.4–725.8 Mbp) on chromosome 3A. For TKW, five novel putative QTLs, i.e., *q.gwas.18* (58.8–68.0 Mbp) on chromosome 3B, *q.gwas.19* (803.2 – 812.9 Mbp) on chromosome 3B, *q.gwas.22* (431.8–442.1 Mbp) on chromosome 5A, *q.gwas.26* (524–534 Mbp) on chromosome 6A, and *q.gwas.34* (681.9–692 Mbp) on chromosome 7A were identified. Fine mapping of these genomic regions is required to identify the genes responsible for these QTLs for SPP and TKW. However, for four of the five novel QTLs for SPP, we were able to identify potential candidate genes, i.e., T*RITD1Bv1G000110* (Tryptophan aminotransferase-related protein 2), *TRITD1Bv1G167110* (UDP-Glucose Pyrophosphorylase 1), *TRITD2Bv1G184650* (ethylene-responsive transcription factor), and *TRITD3Av1G275580* (E3 ubiquitin-protein ligase). *TRITD2Bv1G184650* has been reported to regulate the initiation and development of spikelets in wheat, particularly when the temperature is low (Yu et al., 2021).

Several putative QTLs for SPP are identified here, i.e., *q.gwas.02* (0.3–5.8 Mbp), *q.gwas.14* (94–104.6 Mbp), *q.gwas.20* (721.4–725.8 Mbp), and *q.gwas.30* (2–35 Mbp) were found co-localized with previously reported QTLs for these traits on chromosomes 1B, 2B, 4A, and 7A, respectively (Golabadi et al., 2011; Mengistu et al., 2016; Kidane et al., 2017a; Roncallo et al., 2017; Mangini et al., 2018; Soriano et al., 2018; Li et al., 2019; Rahimi et al., 2019; Alipour et al., 2021). Similarly, putative QTLs for TKW were co-localized with QTLs previously identified, i.e., *q.gwas.08* (MTA for AX-158606713) with a QTL identified based on Ethiopian durum wheat germplasm (Mengistu et al., 2016), and *q.gwas.30* with a QTL identified by Golabadi et al. (2011) based on F3 and F4 populations of durum wheat in Iran, and by Mangini et al. (2018) from a collection of tetraploid wheat grown in Southern Italy. The genomic region regarded as QTL *q.gwas.30* in this study was associated with four traits (DTH, DTM, SPP, TKW) (Table 5). This suggests that either the same gene with pleiotropic effects is involved in regulating these traits, or different genes in this genomic region regulate their corresponding traits or a combination of both. Thus, further research is required to identify common SNP markers representing the four traits in this genomic region and subsequent use in marker-assisted selection for improving the crop. Several genes encoding growth-regulating factor, seed maturation protein, phosphate transporter, phototropic-responsive NPH3 protein G, disease resistance protein RPM1, phosphate-responsive 1 family protein, E3-ubiquitin-protein ligase SINA-like 10, potassium transporter, and chloroplast envelope membrane protein, are among the likely candidates for the QTL q.gwas.30 (Supplementary Table 7).

The marker-trait association analysis conducted *via* GWAS discovered novel and previously identified genomic regions (putative QTLs) associated with DTH. Of these, *q.gwas.01* (522.7–528.6 Mbp of chromosome 1B), *q.gwas.09* (753.4–757.6 Mbp of chromosome 2A), and *q.gwas.30* (2–35 Mbp Mbp of chromosome 3B) were previously reported for this trait (Golabadi et al., 2011; Roncallo et al., 2017; Mangini et al., 2018). These QTLs are significant at two or more test locations and hence can be considered stable MTAs across environments. The present findings also confirmed the results reported in previous studies for DTH on chromosomes 1B, 2A, and 3B (Kidane et al., 2017b; Ogbonnaya et al., 2017; Li et al., 2019; Rahimi et al., 2019; Wang et al., 2019). Hence, there is solid evidence of genes regulating DTH in these genomic regions. One of the novel putative QTLs for DTH is *q.gwas.35*, covering a

151.9–165.3 Mbp region on chromosome 7B. This genomic region contains the *TRITD7Bv1G057630* gene that encodes Flowering Locus T/Terminal Flower 1-like protein. Previous research on wheat, soybean, and Arabidopsis indicates that this gene is located in the region flanking *FT-D1*, a major gene regulating flowering in wheat, soybean, and Arabidopsis (Sun et al., 2019; Isham et al., 2021). Hence, if breeders aim to improve durum wheat for DTH, it is advisable to consider the QTL regions of *q.gwas.35* to get information related to DTH. Furthermore, as previously shown (Wu et al., 2008; Zhao et al., 2020), the *q.gwas.01* QTL region contains a potential candidate gene *TRITD1Bv1G168480* (WRKY). This gene involved in regulating leaf senescence. It is also known to have major roles at various stages of wheat development affecting productivity and product quality and was predicted to interact with DTH. Similarly, the *TRITD7Av1G017240* (zinc finger CCCH-type transcription factor) gene, which promotes wheat flowering, was also identified in this study. Hence, it would be worthwhile to conduct further research on this genomic region to identify the gene involved in *q.gwas.01* and to understand the relationship between leaf senescence and DTH in durum wheat.

A previously known genomic region and three novel genomic regions (putative QTLs) associated with DTM were found on chromosomes 2A, 6A, and 7A. These QTLs explained 2–33% of the variation in DTM. The QTL designated as *q.gwas.25* (601.5–615.3 Mbp of chromosome 6A) was reported in a previous study on wheat (Mangini et al., 2018). The novel putative QTLs are located on chromosomes 2A (67.5–70.6 Mbp; *q.gwas.10*) and 7A (2.1–35.1 Mbp; *q.gwas.30*, and 35.5–68.8 Mbp; *q.gwas.31*). The *TRITD6Av1G220960* (Ethylene receptor) gene, which is located within the genomic region of *q.gwas.25*, is a potential candidate gene for *q.gwas.25*. Previous research has suggested that ethylene receptors are most likely related to the duration of seed development and maturation; i.e., the duration embryo development (Hays et al., 2007). However, in previous findings in maize, grain yield increments were observed through ethylene signal reduction (Shi et al., 2015). Similar to the previous study (Han et al., 2021), *TRITD7Av1G011750* (growth-regulating factor), *TRITD6Av1G017240* (Zinc finger CCCH domain protein TE), and *TRITD7Av1G017550* (NAC domain-containing protein) genes have been found in the genomic regions of QTLs for DTH identified in this study, and are potential candidate genes for the corresponding QTLs. Studies have shown that these genes are mainly involved in regulating growth, development, biotic and abiotic stress adaptation in wheat, rice, and other crop plants and may also determine the variation in phenological traits in wheat and rice.

MTA analysis for PHT identified four putative QTLs, *q.gwas.03* (580.6–585.4 Mbp of chromosomes 1B), *q.gwas.04* (669.5–674.7 Mbp of chromosomes 1B) and *q.gwas24* (628.2–630.9 Mbp of chromosomes 5B), similar to the previous reports (Zhang et al., 2012; Roncallo et al., 2017). This study also found novel MTAs and putative QTLs on chromosome 1B (q.gwas.05; 628.2 – 630.9 Mbp), 5A (q.gwas.22; 431.8–442 Mbp), 6B (q.gwas.29; 119.6–121.5 Mbp), and 7B (q.gwas.37; 686.4–697.9 Mbp). These results show that all significant MTAs for PHT were on the B genome except one MTA on 5A. This may serve as an indicator of potential hotspot regions for genes associated with PHT. The *TRITD1Bv1G191400* (Zink finger protein) could be a candidate gene underlying the *q.gwas.03* QTL for PHT. Previous research revealed that this gene is significantly associated with improved salt tolerance and regulates stress resistance in wheat, Arabidopsis, and other plants (Ciftci-Yilmaz and Mittler, 2008; Ma et al., 2016). Other potential candidate genes for the PHT QTLs include *TRITD1Bv1G189370* (encoding Calcineurin B-like protein), *TRITD1Bv1G189570* (encoding Receptor-like protein kinase), and *TRITD5Bv1G0001780* (Cytochrome P450-like protein).

The present study confirmed previously reported QTL for SPL on chromosome 1B (*q.gwas.06*; 314.3–318.8 Mbp), which was previously reported by Giraldo et al. (2016). Likewise, putative QTL *q.gwas.11* on chromosome 2A (489.5–492.5 Mbp), QTL *q.gwas.12* on chromosome 2A (522.4–534.5 Mbp), QTL *q.gwas.13* on chromosome 2B (185.5–195.6 Mbp), and QTL *q.gwas.23* on chromosome 5A (526.3–534.1Mbp) were found for this trait. These are likely to be novel QTLs since the corresponding genomic regions are not associated with SPL in previous studies. The potential candidate genes for the SPL QTLs include *TRITD2Av1G189490* (*encoding* Acyl-CoA N-acyltransferase), *TRITD2Av1G190600* (encoding Ring finger protein), and *TRITD2Bv1G109560* (encoding E3 ubiquitin-protein ligase). A report from a previous study revealed that E3 ubiquitin proteins have a potential role in modulating crop productivity by influencing growth, development, and important agronomic traits (Varshney and Majee, 2022).

5 Conclusions

In the present study, we evaluated the diverse germplasm of Ethiopian durum wheat using multi-environment trials (MET) data that are genotyped with the Illumina Infinium 25k wheat SNP array to unravel genomic regions associated with its phenological and plant architecture traits as well as grain yield and yield related traits using GWAS. The GWAS identified 44 significant MTAs, including 26 novel genomic regions. The combined analysis of variance revealed significant effects of genotype, environment, and genotype-by-environment interaction on the target traits. The study also confirmed several previously reported QTLs. The identification of a large number of novel QTLs in this study indicates the presence of novel alleles of the genes underlying these QTLs, which probably confirms the distinctness of the Ethiopian durum wheat gene pool from other durum wheat gene pools. The major significant QTLs, such as *q.gwas.21* (for SNP IAAV3365, stable across location) that explained 44.95% of the variation in grain yield, *q.gwas.10* (for SNP Kukri_rep_c73477_888)

that explained 32.9% of the variation in days to maturity and *q.gwas.28* (for SNP Kukri_c60966_261) that explained 30.7% of the variation in the grain-filling period are the key findings of this study. Additional research is needed to validate these key findings, including fine mapping to determine the underlying genes and their subsequent functional analysis. The addition of SNP markers associated with the target traits of this study is highly beneficial for genomic-led breeding of durum wheat.

The results could empower the sustainability of durum breeding by unlocking genomic regions governing complex plant characteristics. Most importantly, the results obtained in the present study could contribute a major role in understanding the durum wheat genome and improving genetic resources for breeding this crop, which in turn, supports global food security. The newly identified genes will also advance the understanding of genomic regions associated with essential characteristics used in durum wheat breeding. The identified novel variants suggest a potential use of Ethiopian durum wheat in durum wheat marker-assisted breeding. The study also provided new insight into the genetic architecture of grain yield and related traits. It indicated the potential of the diverse Ethiopian durum wheat gene pool for future improvement programs. Hence, the identified MTAs and candidate genes could be used to understand the genetic basis of genomic regions of important traits and to accelerate the development of new cultivars with high grain yield and agronomically essential traits *via* precision breeding. In addition, the identified MTAs could be used in marker-assisted breeding, fine mapping, and cloning of the underlying genomic regions and putative QTLs in durum wheat germplasm.

Data availability statement

The genotypic data presented in this study were generated using a commercially available Illumina Infinium 25k wheat SNP array whose details can be found at https://www.traitgenetics.com/index.php/service-products. The genotypic data of the 420 genotypes studied can be obtained upon request.

Author contributions

Conceptualization: BM, KT, RO, MG, EJ, TH, Methodology: BM, MG, KT, RO, EJ, TH, CH, Data curation: BM, Formal analysis: BM, Visualization: BM, MG, Investigation: BM, MG, RO, EJ, Resources: KT, RO, EJ, MG, TH, Funding acquisition: KT, RO, EJ, MG, TH, Project administration: KT, RO, MG, EJ, TH, Supervision: KT, RO, EJ, MG, TH, CH, FH, Writing original draft: BM, Writing-review, and editing: BM, RO, EJ, MG, KT, TH, CH, and FH. All authors have read and approved the final version of the manuscript.

Funding

This work was funded by the Swedish International Development Cooperation Agency (Sida) grant awarded to Addis Ababa University and the Swedish University of Agricultural Sciences for a bilateral capacity-building program in biotechnology. The funding information is available on "https://sida.aau.edu.et/index.php/biotechnology-phdprogram/;accessed on May 21, 2022".The funders played no role in the design of the study, data collection, analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors would like to thank the Institute of Biotechnology, Addis Ababa University (AAU) for the technical support received during the fieldwork and the Swedish University of Agricultural Science for different facility support during seedling development in the greenhouse. The authors are grateful to Sinana Agricultural Research Center (SARC), Kulumsa Agricultural Research Center (KARC), Debrezeit Agricultural Research Center (DZARC), and Wolmera Agricultural Office for providing their experimental sites and equipment for the field trials. The unreserved support from the technical staff of SARC, KARC, and DZARC during the field trial is highly appreciated. We would also like to thank Ethiopian Biodiversity Institute, Wollo University, SARC, and DZARC for providing durum wheat germplasm.

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.2022.1009244/full#supplementary-material

SUPPLEMENTARY TABLE 2
Description of mean monthly weather information of test sites during crop growing season.

SUPPLEMENTARY TABLE 5
Computed physical distance at which mean linkage disequilibrium decay drops below at cut off (r^2 = 0.2) value on pairwise comparison of SNP markers.

SUPPLEMENTARY TABLE 6
GWAS complete output for phenological, plant architecture, grain yield, and yield-related traits across five test sites and combined data. "Trait" indicate measured phenotypic data including; DTH, Days to heading; DTM, Days to physiological maturity; NET, Number of effective tillers per plant; GFP, Grain filling period; PHT, Plant height; SPL, Spike length; SPP, Number of spikelets per spike; TKW, Thousand kernel weight; and GYD, Grain yield. Loc, test sites, AK, CD, HO, KU, and SN are Akaki, Chefe Donsa, Holeta, Kulumsa, and Sinana test sites, respectively. CO; reports combined environments data; CHR, chromosome; SNP, Single nucleotide polymorphisms; POS, Physical position; MAF, minor allele frequency; P.value, reports the significance of the nominal tests and PVE, reports phenotypic variance explained. This Supplementary Table is also presented in Excel format.

SUPPLEMENTARY TABLE 7
High-confidence candidate genes were identified through marker-trait association analysis to associate with phenological, plant architecture, grain yield, and yield-related traits collected at five test sites. "Trait" indicates measured phenotypic data including; DTH, Days to heading; DTM, Days to physiological maturity; NET, Number of effective tillers per plant; GFP, Grain filling period; PHT, Plant height; SPL, Spike length; SPP, Number of spikelets per spike; TKW, Thousand kernel weight; and GYD, Grain yield. MTAs, Marker trait associations; SNPs, Single nucleotide polymorphisms. This Supplementary Table is also presented in Excel format.

SUPPLEMENTARY FIGURE 1
Genome-wide LD decay plot over total physical distance based on 10,045 SNP markers. The yellow curve represents the model that fits LD decay. The solid red line represents the arbitrary threshold for no LD used (r^2 = 0.3). The light green line indicates the intersection between the critical and the map distance to determine QTL confidence intervals.

SUPPLEMENTARY FIGURE 2
The circular Manhattan and Q-Q plots of GWAS results for DTH on panels **(A, B)**, and DTM on panels **(C, D)** were produced using each test site and combined data across test sites. The circular Manhattan plots represent the relative positions of the SNP markers on each chromosome in a circular manner. To view the significant MTAs results, move from outside to the center of each circle, rotating (rounding) through each circle starting from FarmCPU.CO.DTH, followed by FarmCPU.SN.DTH, FarmCPU.KU.DTH, FarmCPU.HO.DTH, FarmCPU.CD.DTH, and FarmCPU.AK.DTH to the center of the circle in sequential order and follow a similar approach for DTM. The name of test sites as stated here is presented in the center of circle. For the QQ−plots, Y-axis represents observed−log10 (p-value), and the X-axis represents expected−log10 (p-value) under the assumption that the *p*-values follow a normal distribution. The dotted lines indicate the 95% confidence interval assuming the null hypothesis of no association between the SNP and trait. DTH refers to days to heading, and DTM refers to days to physiological maturity. CO; Combined data across five environments, SN: Sinana site, KU: Kulumsa site, HO; Holeta site, CD: Chefe Donsa site, and AK: Akaki sites. The Circular Manhattan and Q-Q plots of GWAS results for GYD on panel **(E, F)**, TKW on panel **(G, H)**, and SPP on panel **(I, J)** were plotted using data from each test site and combined data from all test sites, respectively. For the circular manhattan plot, follow a similar approach for these traits as in Figure S2A for all test sites and traits. The assumption of QQ−plots in also applies here. GYD, Grain yield, TKW, Thousand-kernel weight, and SPP, number of spikelets per spike. CO; Combined data across five environments, SN: Sinana site, KU: Kulumsa site, HO; Holeta site, CD: Chefe Donsa site, and AK: Akaki sites. The Circular Manhattan and Q-Q plots of GWAS results for PHT on panel **(K, L)**, NET on panel **(M, N)**, and SPL on panel **(O, P)** were plotted using data from each test site and combined data from all test sites, respectively. For the circular manhattan plot, follow a similar approach for these traits as in Figure S 2A for all test sites and traits. For the QQ−plots, apply the assumption in a. SPL, Spike length; NET, Number of effective tillers per plant; and PHT, Plant height, and SPP. CO; Combined data across five environments, SN, Sinana site; KU, Kulumsa site; HO, Holeta site; CD, Chefe Donsa site; and AK, Akaki sites.

References

Abu-Zaitoun, S. Y., Chandrasekhar, K., Assili, S., Shtaya, M. J., Jamous, R. M., Mallah, O. B., et al. (2018). Unlocking the genetic diversity within a middle-east panel of durum wheat landraces for adaptation to semi-arid climate. *Agronomy* 8 (10), 233. doi: 10.3390/agronomy8100233

Alaux, M., Rogers, J., Letellier, T., Flores, R., Alfama, F., Pommier, C., et al. (2018). Linking the international wheat genome sequencing consortium bread wheat reference genome sequence to wheat genetic and phenomic data. *Genome Biol.* 19(1), 111. doi: 10.1186/s13059-018-1491-4

Alemu, A., Brazauskas, G., Gaikpa, D. S., Henriksson, T., Islamov, B., Jørgensen, L. N., et al. (2021a). Genome-wide association analysis and genomic prediction for adult-plant resistance to *Septoria tritici* blotch and powdery mildew in winter wheat. *Front. Genet.* doi: 10.3389/fgene.2021.661742

Alemu, A., Feyissa, T., Letta, T., and Abeyo, B. (2020a). Genetic diversity and population structure analysis based on the high density SNP markers in Ethiopian durum wheat (*Triticum turgidum* ssp. durum). *BMC Genet.* 21(1), 18. doi: 10.1186/s12863-020-0825-x

Alemu, A., Feyissa, T., Maccaferri, M., Sciara, G., Tuberosa, R., Ammar, K., et al. (2021b). Genome-wide association analysis unveils novel QTLs for seminal root system architecture traits in Ethiopian durum wheat. *BMC Genomics.* 22, 20. doi: 10.1186/s12864-020-07320-4

Alemu, A., Feyissa, T., Tuberosa, R., Maccaferri, M., Sciara, G., Letta, T., et al. (2020b). ScienceDirect genome-wide association mapping for grain shape and color traits in Ethiopian durum wheat (triticum turgidum ssp . durum). *Crop J.* 8 (5), 757–768. doi: 10.1016/j.cj.2020.01.001

Alemu, S. K., Huluka, A. B., Tesfayei, K., Haileselassie, T., and Uauy, C. (2021). Genome-wide association mapping identifies yellow rust resistance loci in Ethiopian durum wheat germplasm. *PloS One.* 16(5), e0243675. doi: 10.1371/journal.pone.0243675

Alipour, H., Abdi, H., Rahimi, Y., and Bihamta, M. R. (2021). Dissection of the genetic basis of genotype-by-environment interactions for grain yield and main agronomic traits in Iranian bread wheat land-races and cultivars. *Sci Rep.* 11 (1):17742. doi: 10.1038/s41598-021-96576-1.

Al-Khayri, J. M., Jain, S. M., and Johnson, D. V. (2016). Advances in plant breeding strategies: Breeding, biotechnology and molecular tools. Switzerland, Springer International Publishing, 656 p. doi: 10.1007/978-3-319-22521-0

Alvarado, G., Rodríguez, F. M., Pacheco, A., Burgueño, J., Crossa, J., Vargas, M., et al. (2020). META-r: A software to analyze data from multi-environment plant breeding trials. *Crop J.* 8(5), 745–756. doi: 10.1016/j.cj.2020.03.010

Anuarbek, S., Abugalieva, S., Pecchioni, N., Laidò, G., Maccaferri, M., Tuberosa, R., et al. (2020). Quantitative trait loci for agronomic traits in tetraploid wheat for

enhancing grain yield in Kazakhstan environments. *PloS One*. 5(6), e0234863. doi: 10.1371/journal.pone.0234863

Aoun, M., Breiland, M., Kathryn Turner, M., Loladze, A., Chao, S., Xu, S. S., et al. (2016). Genome-wide association mapping of leaf rust response in a durum wheat worldwide germplasm collection. *Plant Genome*. 9(3). doi: 10.3835/plantgenome2016.01.0008

Aoun, M., Rouse, M. N., Kolmer, J. A., Kumar, A., and Elias, E. M. (2021). Genome-wide association studies reveal all-stage rust resistance loci in elite durum wheat genotypes. *Front. Plant Sci.* 12, 640739. doi: 10.3389/fpls.2021.640739

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software*. doi: 10.18637/jss.v067.i01

Bellucci, A., Tondelli, A., Fangel, J. U., Torp, A. M., Xu, X., Willats, W. G. T., et al. (2017). Genome-wide association mapping in winter barley for grain yield and culm cell wall polymer content using the high-throughput CoMPP technique. *PloS One*. 12(3), e0173313. doi: 10.1371/journal.pone.0173313

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc Ser. B.* 57(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bhatta, M., Morgounov, A., Belamkar, V., and Baenziger, P. S. (2018). Genome-wide association study reveals novel genomic regions for grain yield and yield-related traits in drought-stressed synthetic hexaploid wheat. *Int. J. Mol. Sci.* 19(10), 3237. doi: 10.3390/ijms19103011

Borrego-Benjumea, A., Carter, A., Zhu, M., Tucker, J. R., Zhou, M., and Badea, A. (2021). Genome-wide association study of waterlogging tolerance in barley (Hordeum vulgare l.) under controlled field conditions. *Front. Plant Sci.* 12, 711654. doi: 10.3389/fpls.2021.711654

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

Breseghello, F., and Sorrells, M. E. (2006a). Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci*. 46(3), 1323–1330. doi: 10.2135/cropsci2005.09-0305

Breseghello, F., and Sorrells, M. E. (2006b). 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

Canè, M. A., Maccaferri, M., Nazemi, G., Salvi, S., Francia, R., Colalongo, C., et al. (2014). Association mapping for root architectural traits in durum wheat seedlings as related to agronomic performance. *Mol. Breed.* 34(4), 1629–1645. doi: 10.1007/s11032-014-0177-1

Ceglar, A., Toreti, A., Zampieri, M., and Royo, C. (2021). Global loss of climatically suitable areas for durum wheat growth in the future. *Environ. Res. Lett*. 16, 104049. doi: 10.1088/1748-9326/ac2d68

Chen, J., Zhang, F., Zhao, C., Lv, G., Sun, C., Pan, Y., et al. (2019). Genome-wide association study of six quality traits reveals the association of the TaRPP13L1 gene with flour colour in Chinese bread wheat. *Plant Biotechnol. J.* 17(11), 2106–2122. doi: 10.1111/pbi.13126

Ciftci-Yilmaz, S., and Mittler, R. (2008). The zinc finger network of plants. *Cell. Mol. Life Sci*. 65, 1150–1160. doi: 10.1007/s00018-007-7473-4

Colasuonno, P., Lozito, M. L., Marcotuli, I., Nigro, D., Giancaspro, A., Mangini, G., et al. (2017). The carotenoid biosynthetic and catabolic genes in wheat and their association with yellow pigments. *BMC Genomics*. 18(1), 122. doi: 10.1186/s12864-016-3395-6

Collard, B. C. Y., and Mackill, D. J. (2008). Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philos. Trans. R. Soc B Biol. Sci.* 363(1491), 557–572. doi: 10.1098/rstb.2007.2170

Condorell, G. E., Newcomb, M., Groli, E. L., Maccaferri, M., Forestan, C., Babaeian, E., et al. (2022). Genome wide association study uncovers the QTLome for osmotic adjustment and related drought adaptive traits in durum wheat. *Genes (Basel)*. 13(2), 293. doi: 10.3390/genes13020293

Corbeil, R. R., and Searle, S. R. (1976). Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*. 18, 31–38. doi: 10.1080/00401706.1976.10489397

Earl, D. A., and vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the evanno method. *Conserv. Genet. Resour*. 4, 359–361. doi: 10.1007/s12686-011-9548-7

Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.*14 (8), 2611–20. doi: 10.1111/j.1365-294X.2005.02553.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(4), 574–578. doi: 10.1111/j.1471-8286.2007.01758.x

Fayaz, F., Aghaee Sarbarzeh, M., Talebi, R., and Azadi, A. (2019). Genetic diversity and molecular characterization of Iranian durum wheat landraces (*Triticum turgidum* durum (Desf.) husn.) using DArT markers. *Biochem. Genet*. 57(1), 98–116. doi: 10.1007/s10528-018-9877-2

Fiedler, J. D., Salsman, E., Liu, Y., Michalak de Jiménez, M., Hegstad, J. B., Chen, B., et al. (2017). Genome-wide association and prediction of grain and semolina quality traits in durum wheat breeding populations. *Plant Genome*. 10(3). doi: 10.3835/plantgenome2017.05.0038

Gao, L., Meng, C., Yi, T., Xu, K., Cao, H., Zhang, S., et al. (2021). Genome-wide association study reveals the genetic basis of yield- and quality-related traits in wheat. *BMC Plant Biol.* 21(1), 144. doi: 10.1186/s12870-021-02925-7

Garcia, M., Eckermann, P., Haefele, S., Satija, S., Sznajder, B., Timmins, A., et al. (2019). Genome-wide association mapping of grain yield in a diverse collection of spring wheat (*Triticum aestivum* l.) evaluated in southern Australia. *PloS One*. 14 (2), e0211730. doi: 10.1371/journal.pone.0211730

Geleta, M., and Ortiz, R. (2016). Molecular and Genomic Tools Provide Insights on Crop Domestication and Evolution. In: L. S. Donald, Editor(s). *Advances in Agronomy*, Academic Press 135, 185–223. doi: 10.1016/bs.agron.2015.09.005.

Ghavami, F., Elias, E. M., Mamidi, S., Mergoum, M., Kianian, S. F., Ansari, O., et al. (2011). Mixed model association mapping for fusarium head blight resistance in tunisian-derived durum wheat populations. *G3 Genes Genomes Genet*. 1(3), 209–218. doi: 10.1534/g3.111.000489

Giraldo, P., Royo, C., González, M., Carrillo, J. M., and Ruiz, M. (2016). Genetic diversity and association mapping for agromorphological and grain quality traits of a structured collection of durum wheat landraces including subsp. durum, turgidum and diccocon. *PloS One*. 11(11), e0166577. doi: 10.1371/journal.pone.0166577

Golabadi, M., Arzani, A., Mirmohammadi Maibody, S. A. M., Tabatabaei, B. E. S., and Mohammadi, S. A. (2011). Identification of microsatellite markers linked with yield components under drought stress at terminal growth stages in durum wheat. *Euphytica*. 177, 207–221. doi: 10.1007/s10681-010-0242-8

Gonçalves-Vidigal, M. C., Mora, F., Bignotto, T. S., Munhoz, R. E. F., and De Souza, L. D. (2008). Heritability of quantitative traits in segregating common bean families using a Bayesian approach. *Euphytica* 164, 551–560. doi: 10.1007/s10681-008-9758-6

Han, G., Qiao, Z., Li, Y., Wang, C., and Wang, B. (2021). The roles of ccch zinc-finger proteins in plant abiotic stress tolerance. *Int. J. Mol. Sci*. 22(15), 8327. doi: 10.3390/ijms22158327

Hays, D. B., Do, J. H., Mason, R. E., Morgan, G., and Finlayson, S. A. (2007). Heat stress induced ethylene production in developing wheat grains induces kernel abortion and increased maturation in a susceptible cultivar. *Plant Sci*. 172, 1113–1123. doi: 10.1016/J.PLANTSCI.2007.03.004

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, 961–967. doi: 10.1038/ng.695.

IBPGR (1985). *Descriptors wheat*, Vol. 12.

Isham, K., Wang, R., Zhao, W., Wheeler, J., Klassen, N., Akhunov, E., et al. (2021). QTL mapping for grain yield and three yield components in a population derived from two high-yielding spring wheat cultivars. *Theor. Appl. Genet*. 134(7), 2079–2095. doi: 10.1007/s00122-021-03806-1

Johansson, E., Branlard, G., Cuniberti, M., Flagella, Z., Hüsken, A., Nurit, E., et al. (2020a). "Genotypic and environmental effects on wheat technological and nutritional quality," in *Wheat Quality For Improving Processing And Human Health*. ed. T. M. Igrejas, C. Ikeda and G. Guzmán (Springer Cham, Springer Nature Switzerland AG), 171–204. doi: 10.1007/978-3-030-34163-3_8

Johansson, E., Henriksson, T., Prieto-Linde, M. L., Andersson, S., Ashraf, R., and Rahmatov, M. (2020b). Diverse wheat-alien introgression lines as a basis for durable resistance and quality characteristics in bread wheat. *Front. Plant Sci*. 11, 1067. doi: 10.3389/fpls.2020.01067

Johansson, E., Prieto-Linde, M. L., and Larsson, H. (2021). Locally adapted and organically grown landrace and ancient spring cereals–a unique source of minerals in the human diet. *Foods*. 10(2), 393. doi: 10.3390/foods10020393

Kabbaj, H., Sall, A. T., Al-Abdallat, A., Geleta, M., Amri, A., Filali-Maltouf, A., et al. (2017). Genetic diversity within a global panel of durum wheat (*Triticum durum*) landraces and modern germplasm reveals the history of alleles exchange. *Front. Plant Sci*. 8, 1277. doi: 10.3389/fpls.2017.01277

Kankwatsa, P., Singh, D., Thomson, P. C., Babiker, E. M., Bonman, J. M., Newcomb, M., et al. (2017). Characterization and genome-wide association mapping of resistance to leaf rust, stem rust and stripe rust in a geographically diverse collection of spring wheat landraces. *Mol. Breed*. 37(3), 113. doi: 10.1007/s11032-017-0707-8

Kidane, Y. G., Gesesse, C. A., Hailemariam, B. N., Desta, E. A., Mengistu, D. K., Fadda, C., et al. (2019). A large nested association mapping population for breeding and quantitative trait locus mapping in Ethiopian durum wheat. *Plant Biotechnol. J.* 17(7), 1380–1393. doi: 10.1111/pbi.13062

Kidane, Y. G., Hailemariam, B. N., Mengistu, D. K., Fadda, C., Pè, M. E., and Dell'Acqua, M. (2017a). Genome-wide association study of septoria tritici blotch

resistance in Ethiopian durum wheat landraces. *Front. Plant Sci.* 8, 1586. doi: 10.3389/fpls.2017.01586

Kidane, Y. G., Mancini, C., Mengistu, D. K., Frascaroli, E., Fadda, C., Pè, M. E., et al. (2017b). Genome wide association study to identify the genetic base of smallholder farmer preferences of durum wheat traits. *Front. Plant Sci.* 8, 1230. doi: 10.3389/fpls.2017.01230

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(3), 1179–1191. doi: 10.1111/1755-0998.12387

Kumar, D., Kumar, A., Chhokar, V., Gangwar, O. P., Bhardwaj, S. C., Sivasamy, M., et al. (2020). Genome-wide association studies in diverse spring wheat panel for stripe, stem, and leaf rust resistance. *Front. Plant Sci.* 11, 748. doi: 10.3389/fpls.2020.00748

Letta, T., Olivera, P., Maccaferri, M., Jin, Y., Ammar, K., Badebo, A., et al. (2014). Association mapping reveals novel stem rust resistance loci in durum wheat at the seedling stage. *Plant Genome.* 7, 1. doi: 10.3835/plantgenome2013.08.0026

Levene, H. (1960). Robust tests for equality of variances. *Contrib. to Probab. Stat. Essays*.

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), 2397–2397. 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

Liu, W., Maccaferri, M., Bulli, P., Rynearson, S., Tuberosa, R., Chen, X., et al. (2017a). Genome-wide association mapping for seedling and field resistance to puccinia striiformis f. sp. tritici in elite durum wheat. *Theor. Appl. Genet.* 130(4), 649–667. doi: 10.1007/s00122-016-2841-9

Liu, W., Maccaferri, M., Chen, X., Laghetti, G., Pignone, D., Pumphrey, M., et al. (2017b). Genome-wide association mapping reveals a rich genetic architecture of stripe rust resistance loci in emmer wheat (Triticum turgidum ssp. dicoccum). *Theor. Appl. Genet.* 130, 2249–2270. doi: 10.1007/s00122-017-2957-6

Liu, W., Maccaferri, M., Rynearson, S., Letta, T., Zegeye, H., Tuberosa, R., et al. (2017c). Novel sources of stripe rust resistance identified by genome-wide association mapping in ethiopian durum wheat (*Triticum turgidum* ssp. durum). *Front. Plant Sci.* 8, 774. doi: 10.3389/fpls.2017.00774

Liu, J., Rasheed, A., He, Z., Imtiaz, M., Arif, A., Mahmood, T., et al. (2019). Genome-wide variation patterns between landraces and cultivars uncover divergent selection during modern wheat breeding. *Theor. Appl. Genet.* 132(9), 2509–2523. doi: 10.1007/s00122-019-03367-4

Li, F., Wen, W., Liu, J., Zhang, Y., Cao, S., He, Z., et al. (2019). Genetic architecture of grain yield in bread wheat based on genome-wide association studies. *BMC Plant Biol.* 19(1), 168. doi: 10.1186/s12870-019-1781-3

Maccaferri, M., Cane, M. A., Sanguineti, M. C., Salvi, S., Colalongo, M. C., Massi, A., et al. (2014). A consensus framework map of durum wheat (Triticum durum desf.) suitable for linkage disequilibrium analysis and genome-wide association mapping. *BMC Genomics*. 15(1), 873. doi: 10.1186/1471-2164-15-873

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

Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., et al. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat. Genet.* 51, 885–895. doi: 10.1038/s41588-019-0381-3

Maccaferri, M., Sanguineti, M. C., Corneti, S., Ortega, J. L. A., Salem, M.B., Bort, J., et al. (2008). Quantitative trait loci for grain yield and adaptation of durum wheat (Triticum durum desf.) across a wide range of water availability. *Genetics.* 178(1), 489–511. doi: 10.1534/genetics.107.077297

Maccaferri, M., Sanguineti, M. C., Mantovani, P., Demontis, A., Massi, A., Ammar, K., et al. (2010). Association mapping of leaf rust response in durum wheat. *Mol. Breed.* 26, 189–228. doi: 10.1007/s11032-009-9353-0

Ma, X., Liang, W., Gu, P., and Huang, Z. (2016). Salt tolerance function of the novel C2H2-type zinc finger protein TaZNF in wheat. *Plant Physiol. Biochem.* 06, 129–140. doi: 10.1016/j.plaphy.2016.04.033

Mangini, G., Gadaleta, A., Colasuonno, P., Marcotuli, I., Signorile, A. M., Simeone, R., et al. (2018). Genetic dissection of the relationships between grain yield components by genome-wide association mapping in a collection of tetraploid wheats. *PloS One.* 13(1), e0190162. doi: 10.1371/journal.pone.0190162

Mathew, I., Shimelis, H., Shayanowako, A. I. T., Laing, M., and Chaplot, V. (2019). Genome-wide association study of drought tolerance and biomass allocation in wheat. *PloS One.* 14(12), e0225383. doi: 10.1371/journal.pone.0225383

Mazzucotelli, E., Sciara, G., Mastrangelo, A. M., Desiderio, F., Xu, S. S., Faris, J., et al. (2020). The global durum wheat panel (GDP): An international platform to identify and exchange beneficial alleles. *Front. Plant Sci.* 11, 569905. doi: 10.3389/fpls.2020.569905

Mekonnen, T., Sneller, C. H., Haileselassie, T., Ziyomo, C., Abeyo, B. G., Goodwin, S. B., et al. (2021). Genome-wide association study reveals novel genetic loci for quantitative resistance to septoria tritici blotch in wheat (*Triticum aestivum* l.). *Front. Plant Sci.* 12, 671323. doi: 10.3389/fpls.2021.671323

Mengistu, K., Catellani, M., Frascaroli, E., Fadda, C., Pè, M. E., and Dell'Acqua, M. (2016). High-density molecular characterization and association mapping in Ethiopian durum wheat landraces reveals high diversity and potential for wheat breeding. *Plant Biotechnol. J.* 14(9), 1800–1812. doi: 10.1111/pbi.12538

Mengistu, D. K., Kiros, A. Y., and Pè, M. E. (2015). Phenotypic diversity in Ethiopian durum wheat (Triticum turgidum var. durum) landraces. *Crop J.* 3(3), 190–199. doi: 10.1016/j.cj.2015.04.003

Mérida-García, R., Bentley, A. R., Gálvez, S., Dorado, G., Solís, I., Ammar, K., et al. (2020). Mapping agronomic and quality traits in elite durum wheat lines under differing water regimes. *Agronomy*. 10(1), 144. doi: 10.3390/agronomy10010144

Mérida-García, R., Liu, G., He, S., Gonzalez-Dugo, V., Dorado, G., Gálvez, S., et al. (2019). Genetic dissection of agronomic and quality traits based on association mapping and genomic selection approaches in durum wheat grown in southern Spain. *PloS One.* 14(2), e0211718. doi: 10.1371/journal.pone.0211718

Mohammadi, R., Armion, M., Zadhasan, E., Ahmadi, M. M., and Amri, A. (2018). The use of AMMI model for interpreting genotype × environment interaction in durum wheat. *Exp. Agric.* 54(5), 670–683. doi: 10.1017/S0014479717000308

Ogbonnaya, F. C., Rasheed, A., Okechukwu, E. C., Jighly, A., Makdis, F., Wuletaw, T., et al. (2017). Genome-wide association study for agronomic and physiological traits in spring wheat evaluated in a range of heat prone environments. *Theor. Appl. Genet.* 130, 1819–1835. doi: 10.1007/s00122-017-2927-z

Ozkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamini, F. (2002). AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *Mol. Biol. Evol.* 19 (10), 1797–801. doi: 10.1093/oxfordjournals.molbev.a004002

Patterson, H. D., and Williams, E. R. (1976). A new class of resolvable incomplete block designs. *Biometrika*. 63(1), 83–92. doi: 10.1093/biomet/63.1.83

Peng, J., Ronin, Y., Fahima, T., Röder, M. S., Li, Y., Nevo, E., et al. (2003). Domestication quantitative trait loci in triticum dicoccoides, the progenitor of wheat. *Proc. Natl. Acad. Sci. U. S. A.* 100(5), 2489–2494. doi: 10.1073/pnas.252763199

Pritchard, J. K., Stephens, P., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics.* 155(2), 945–959. doi: 10.1093/genetics/155.2.945

Quamruzzaman, M., Manik, S. M. N., Shabala, S., Cao, F., and Zhou, M. (2021). Genome-wide association study reveals a genomic region on 5AL for salinity tolerance in wheat. *Theor. Appl. Genet.* 135(2), 709–721. doi: 10.1007/s00122-021-03996-8

Rahimi, Y., Bihamta, M. R., Taleei, A., Alipour, H., and Ingvarsson, P. K. (2019). Genome-wide association study of agronomic traits in bread wheat reveals novel putative alleles for future breeding programs. *BMC Plant Biol.* 19(1), 149. doi: 10.1186/s12870-019-1754-6

Rajaram, S., Van Ginkel, M., and Fischer, R. A. (1994). CIMMYT's wheat breeding mega-environments. *Proceedings of the 8th International Wheat Genetics Symposium*, Beijing, China.

R Development Core team (2021). R: a language and environment for statistical computing. *Version 4.0.5. Vienna, Austria: R Foundation for Statistical Computing*; https://www.R-project.org/ . R. A. Lang. Environ. Stat. Comput. R Found. Stat. Comput. Vienna, Austria. ISBN 3-900051-07-0, URL http//www.R-project.org/

Roncallo, P. F., Akkiraju, P. C., Cervigni, G. L., and Echenique, V. C. (2017). QTL mapping and analysis of epistatic interactions for grain yield and yield-related traits in triticum turgidum l. var. durum. *Euphytica*. 213, 277. doi: 10.1007/s10681-017-2058-2

Roncallo, P. F., Beaufort, V., Larsen, A. O., Dreisigacker, S., and Echenique, V. (2018). Genetic diversity and linkage disequilibrium using SNP (KASP) and AFLP markers in a worldwide durum wheat (*Triticum turgidum* l. var durum) collection. *PloS One.* 14, 1–33. doi: 10.1371/journal.pone.0218562

Rufo, R., Alvaro, F., Royo, C., and Soriano, J. M. (2019). From landraces to improved cultivars: Assessment of genetic diversity and population structure of Mediterranean wheat using SNP markers. *PloS One.* 14(7), e0219867. doi: 10.1371/journal.pone.0219867

Sabadin, P. K., Malosetti, M., Boer, M. P., Tardin, F. D., Santos, F. G., Guimarães, C. T., et al. (2012). Studying the genetic basis of drought tolerance in sorghum by managed stress trials and adjustments for phenological and plant height differences. *Theor. Appl. Genet.* 124(8), 1389–402. doi: 10.1007/s00122-012-1795-9

Sall, A. T., Chiari, T., Legesse, W., Seid-Ahmed, K., Ortiz, R., Van Ginkel, M., et al. (2019). Durum wheat (*Triticum durum* desf.): Origin, cultivation and potential expansion in sub-saharan Africa. *Agronomy*. 9(5), 263. doi: 10.3390/agronomy9050263

Sansaloni, C., Franco, J., Santos, B., Percival-Alwyn, L., Singh, S., Petroli, C., et al. (2020). Diversity analysis of 80,000 wheat accessions reveals consequences and opportunities of selection footprints. *Nat. Commun*. 11(1), 4572. doi: 10.1038/s41467-020-18404-w

Shi, J., Habben, J. E., Archibald, R. L., Drummond, B. J., Chamberlin, M. A., Williams, R. W., et al. (2015). Overexpression of ARGOS genes modifies plant sensitivity to ethylene, leading to improved drought tolerance in both arabidopsis and maize. *Plant Physiol.* 169(1), 266–82. doi: 10.1104/pp.15.00780

Siol, M., Jacquin, F., Chabert-Martinello, M., Smýkal, P., Le Paslier, M. C., Aubert, G., et al. (2017). Patterns of genetic structure and linkage disequilibrium in a large collection of pea germplasm. *G3 Genes Genomes Genet*. 7(8), 2461–2471. doi: 10.1534/g3.117.043471

Soriano, J. M., Villegas, D., Aranzana, M. J., García Del Moral, L. F., and Royo, C. (2016). Genetic structure of modern durum wheat cultivars and mediterranean landraces matches with their agronomic performance. *PloS One*. 11(8), e0160983. doi: 10.1371/journal.pone.0160983

Soriano, J. M., Villegas, D., Sorrells, M. E., and Royo, C. (2018). Durum wheat landraces from east and west regions of the mediterranean basin are genetically distinct for yield components and phenology. *Front. Plant Sci*. 8(80), 40. doi: 10.3389/fpls.2018.00080

Sukumaran, S., Reynolds, M. P., and Sansaloni, C. (2018). Genome-wide association analyses identify QTL hotspots for yield and component traits in durum wheat grown under yield potential, drought, and heat stress environments. *Front. Plant Sci*. 9, 81. doi: 10.3389/fpls.2018.00081

Sun, F., Xu, M., Park, C., Dwiyanti, M. S., Nagano, A. J., Zhu, J., et al. (2019). Characterization and quantitative trait locus mapping of late-flowering from a Thai soybean cultivar introduced into a photoperiod-insensitive genetic background. *PloS One*. 14(12), e0226116. doi: 10.1371/journal.pone.0226116

Suprayogi, Y., Pozniak, C. J., Clarke, F. R., Clarke, J. M., Knox, R. E., and Singh, A. K. (2009). Identification and validation of quantitative trait loci for grain protein concentration in adapted Canadian durum wheat populations. *Theor. Appl. Genet*. 119(3), 437–48. doi: 10.1007/s00122-009-1050-1

Talini, R. F., Brandolini, A., Miculan, M., Brunazzi, A., Vaccino, P., Mario Enrico, P., et al. (2020). Genome-wide association study of agronomic and quality traits in a world collection of the wild wheat relative triticum urartu. *Plant J*. 102(3), 555–568. doi: 10.1111/tpj.14650

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., et al. (2016). GAPIT version 2: An enhanced integrated tool for genomic association and prediction. *Plant Genome*. 9(2). doi: 10.3835/plantgenome2015.11.0120

Tuberosa, R. (2012). Phenotyping for drought tolerance of crops in the genomics era. *Front. Physiol*. 3, 347. doi: 10.3389/fphys.2012.00347

Turki, N., Shehzad, T., Harrabi, M., and Okuno, K. (2015). Detection of QTLs associated with salinity tolerance in durum wheat based on association analysis. *Euphytica*. 201, 29–41. doi: 10.1007/s10681-014-1164-7

Tzarfati, R., Barak, V., Krugman, T., Fahima, T., Abbo, S., Saranga, Y., et al. (2014). Novel quantitative trait loci underlying major domestication traits in tetraploid wheat. *Mol. Breed*. 34, 1613–1628. doi: 10.1007/s11032-014-0182-4

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. 91(11), 4414–4423. doi: 10.3168/jds.2007-0980

Varshney, V., and Majee, M. (2022). Emerging roles of the ubiquitin – proteasome pathway in enhancing crop yield by optimizing seed agronomic traits. *Plant Cell Rep*. 41(9), 1805–1826. doi: 10.1007/s00299-022-02884-9

Velu, G., Tutus, Y., Gomez-Becerra, H. F., Hao, Y., Demir, L., Kara, R., et al. (2017). QTL mapping for grain zinc and iron concentrations and zinc efficiency in a tetraploid and hexaploid wheat mapping populations. *Plant Soil*. 8, 680391. doi: 10.1007/s11104-016-3025-8

Walter, S., Kahla, A., Arunachalam, C., Perochon, A., Khan, M. R., Scofield, S. R., et al. (2015). A wheat ABC transporter contributes to both grain formation and mycotoxin tolerance. *J. Exp. Bot*. 66(9), 2583–2593. doi: 10.1093/jxb/erv048

Wang, S., Xu, S., Chao, S., Sun, Q., Liu, S., and Xia, G. (2019). A genome-wide association study of highly heritable agronomic traits in durum wheat. *Front. Plant Sci*. 10, 919. doi: 10.3389/fpls.2019.00919

Wang, J., and Zhang, Z. (2021). GAPIT version 3: Boosting power and accuracy for genomic association and prediction. *Genomics Proteomics Bioinf*. 19(4), 629–640. doi: 10.1016/j.gpb.2021.08.005

Wanke, D., and Üner Kolukisaoglu, H. (2010). An update on the ABCC transporter family in plants: Many genes, many proteins, but how many functions? *Plant Biol*. 12(1), 15–25. doi: 10.1111/j.1438-8677.2010.00380.x

Weir, B. S. (1997). Genetic data analysis II. *Biometrics*. doi: 10.2307/2533134

Wu, H., Ni, Z., Yao, Y., Guo, G., and Sun, Q. (2008). Cloning and expression profiles of 15 genes encoding WRKY transcription factor in wheat (Triticum aestivem l.). *Prog. Nat. Sci*. 18(6), 697–705. doi: 10.1016/j.pnsc.2007.12.006

Xynias, I. N., Mylonas, I., Korpetis, E. G., Ninou, E., Tsaballa, A., Avdikos, I. D., et al. (2020). Durum wheat breeding in the Mediterranean region: Current status and future prospects. *Agronomy*.10(3), 432. doi: 10.3390/agronomy10030432

Yu, X., Jiang, Y., Yao, H., Ran, L., Zang, Y., and Xiong, F. (2021). Cytological and molecular characteristics of delayed spike development in wheat under low temperature in early spring. *Crop J*.10(3), 840–852. doi: 10.1016/j.cj.2021.08.008

Zhang, L., Luo, J. T., Hao, M., Zhang, L. Q., Yuan, Z. W., Yan, Z. H., et al. (2012). Genetic map of triticum turgidum based on a hexaploid wheat population without genetic recombination for d genome. *BMC Genet*. 13, 69. doi: 10.1186/1471-2156-13-69

Zhao, M. M., Zhang, X. W., Liu, Y. W., Li, K., Tan, Q., Zhou, S., et al. (2020). A WRKY transcription factor, TaWRKY42-b, facilitates initiation of leaf senescence by promoting jasmonic acid biosynthesis. *BMC Plant Biol*. 20(1), 444. doi: 10.1186/s12870-020-02650-7

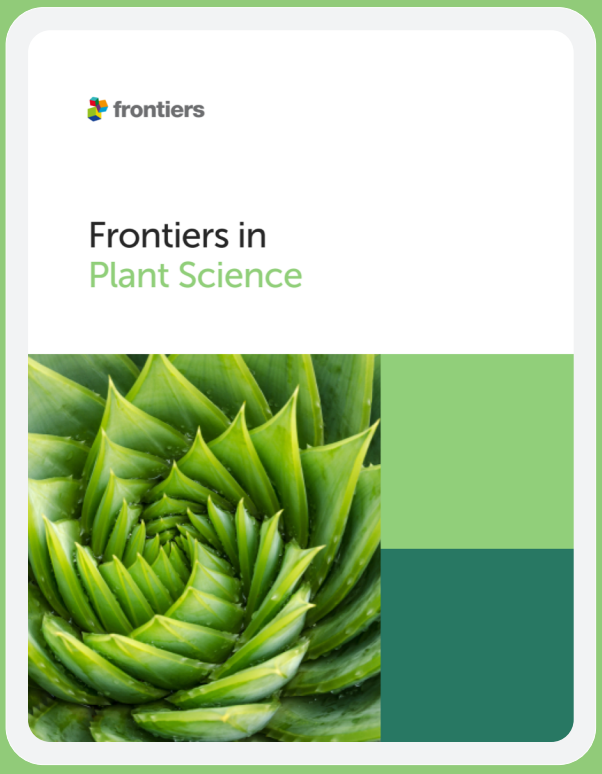
Frontiers in Plant Science

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