

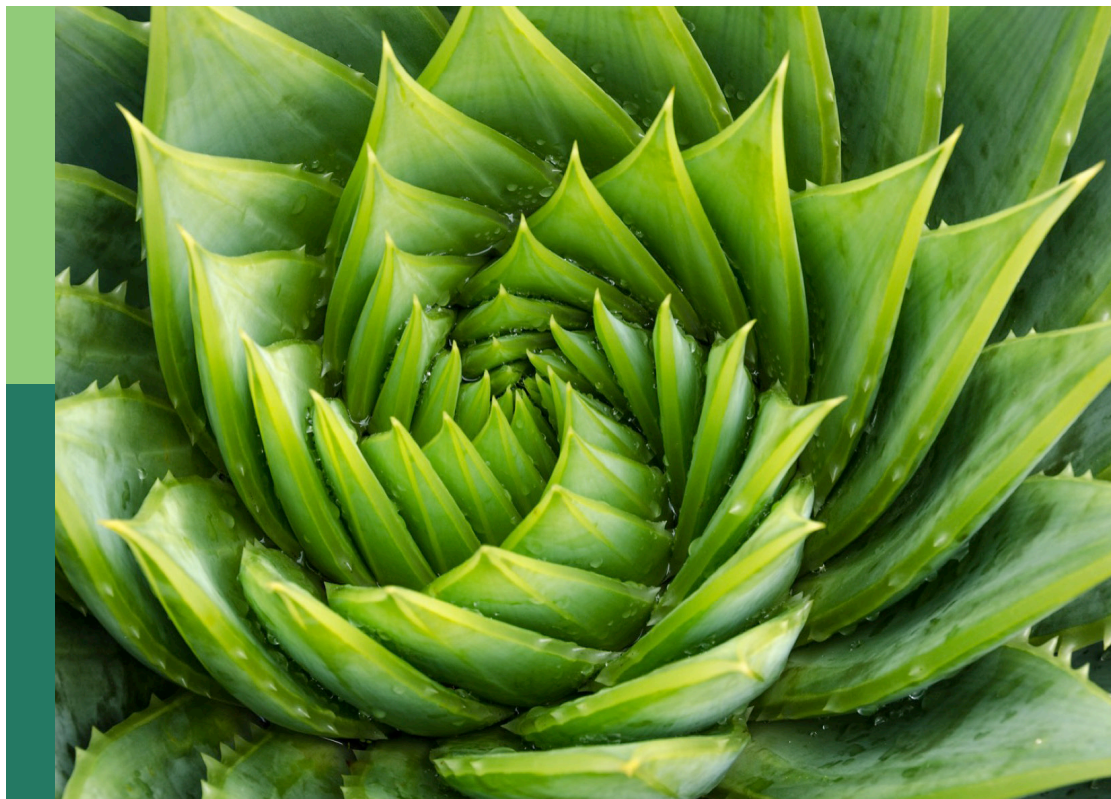
Genomics-driven advances in crop productivity and stress resilience

Edited by

Parimalan Rangan, Robert Henry and
Ambika Baldev Gaikwad

Published in

Frontiers in Plant Science



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-6641-1
DOI 10.3389/978-2-8325-6641-1

Generative AI statement

Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Genomics-driven advances in crop productivity and stress resilience

Topic editors

Parimalan Rangan — National Bureau of Plant Genetic Resources, Indian Council of Agricultural Research (ICAR), India

Robert Henry — The University of Queensland, Australia

Ambika Baldev Gaikwad — National Bureau of Plant Genetic Resources, Indian Council of Agricultural Research (ICAR), India

Citation

Rangan, P., Henry, R., Gaikwad, A. B., eds. (2025). *Genomics-driven advances in crop productivity and stress resilience*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-6641-1

Table of contents

- 05 **Editorial: Genomics-driven advances in crop productivity and stress resilience**
Parimalan Rangan, Robert Henry and Ambika Baldev Gaikwad
- 08 **Identification of QTL-allele systems of seed size and oil content for simultaneous genomic improvement in Northeast China soybeans**
Jianbo He, Lianshun Fu, Xiaoshuai Hao, Yicun Wu, Mengfan Wang, Qi Zhang, Weidan Feng, Mengmeng Fu, Yanping Wang, Haixiang Ren, Weiguang Du, Wubin Wang and Junyi Gai
- 22 **Mapping heat tolerance QTLs in *Triticum durum*-*Aegilops speltoides* backcross introgression lines to enhance thermotolerance in wheat**
Navaneetha Krishnan J., Satinder Kaur, Uttam Kumar, Rohtas Singh, Guriqbal Singh Dhillon, Pradeep Kumar Bhati and Parveen Chhuneja
- 39 **Development of KASP markers, SNP fingerprinting and population genetic analysis of *Cymbidium ensifolium* (L.) Sw. germplasm resources in China**
Baoming Shen, Airong Shen, Yun Tan, Lina Liu, Sainan Li and Zhuming Tan
- 54 **History and prospects of flax genetic markers**
Daiana A. Zhernova, Elena N. Pushkova, Tatiana A. Rozhmina, Elena V. Borkhert, Alexander A. Arkhipov, Elizaveta A. Sigova, Ekaterina M. Dvorianinova, Alexey A. Dmitriev and Nataliya V. Melnikova
- 63 **Harnessing the diversity of a lettuce wild relative to identify anthocyanin-related genes transcriptionally responsive to drought stress**
Inés Medina-Lozano, Jérôme Grimplet and Aurora Díaz
- 81 **Combining genotyping approaches improves resolution for association mapping: a case study in tropical maize under water stress conditions**
Fernanda Carla Ferreira de Pontes, Ingrid Pinheiro Machado, Maria Valnice de Souza Silveira, Antônio Lucas Aguiar Lobo, Felipe Sabadin, Roberto Fritsche-Neto and Júlio César DoVale
- 99 **Phenotypic adaptation and genomic variation of *Kandelia obovata* associated with its northern introduction along southeastern coast of China**
Jiaqi Zhang, Sheng Ouyang, Xingxing Cai, Sheng Yang, Qiuxia Chen, Ji Yang, Zhiping Song, Wenju Zhang, Yuguo Wang, Yan Zhu and Peng Nan

- 116 **Mitochondrial genome assembly and comparative analysis of decaploid *Camellia hainanica***
Shihui Zhang, Yuyan Zhang, Sheng Luo, Jie Gao, Haiyan Hu, Jinping Liu, Wenqiang Wu, Jian Wang, Xiaolong Huang, Hanggui Lai and Dongyi Huang
- 135 **Heat waves reveal additive genetic effects leading to sunburn resilience of grapevine berries**
Tom Heinekamp, Franco Röckel, Katja Herzog, Oliver Trapp, Reinhard Töpfer and Florian Schwander



OPEN ACCESS

EDITED AND REVIEWED BY
Huihui Li,
Chinese Academy of Agricultural Sciences,
China

*CORRESPONDENCE
Parimalan Rangan
✉ r.parimalan@icar.org.in

RECEIVED 12 June 2025
ACCEPTED 30 June 2025
PUBLISHED 11 July 2025

CITATION
Rangan P, Henry R and Gaikwad AB (2025)
Editorial: Genomics-driven advances in
crop productivity and stress resilience.
Front. Plant Sci. 16:1645714.
doi: 10.3389/fpls.2025.1645714

COPYRIGHT
© 2025 Rangan, Henry and Gaikwad. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Editorial: Genomics-driven advances in crop productivity and stress resilience

Parimalan Rangan^{1,2*}, Robert Henry²
and Ambika Baldev Gaikwad¹

¹Division of Genomic Resources, ICAR-National Bureau of Plant Genetic Resources, New Delhi, India,
²Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia,
QLD, Australia

KEYWORDS

genomics, germplasm (genetic) resources, climate change, crop wild relatives (CWRs),
crop productivity, stress resilience

Editorial on the Research Topic

Genomics-driven advances in crop productivity and stress resilience

Scaling up crop productivity in response to climate change is critical to sustainably feeding the ever-growing population. The rate of genetic gain being achieved in recent decades needs to be augmented to meet this demand (van Dijk et al., 2021; Hunter et al., 2017). Genomic selection and gene editing strategies for *de novo* domestication, breaking linkage drag, and overcoming genetic incompatibility barriers are genomics-driven tools that have been demonstrated to enhance crop productivity and stress resilience. Strategies such as landscape genomics, which also consider environmental variables, increase the potential utilization of genebank collections, including crop wild relatives (CWRs) and 'exotic genetic libraries' through identification of appropriate accessions for utilization (Bohra et al., 2022; Campbell et al., 2025; Shrestha et al., 2025). Genomics tools also help gain novel insights into the genetic and epigenetic mechanisms associated with stress resilience traits and the factors to target for enhanced crop productivity (Bailey-Serres et al., 2019; Gupta, 2025; Lohani et al., 2025; Miryeganeh, 2025). Understanding the role of the interplay of various components such as transposable elements (Tossolini et al., 2025) secondary metabolites (Khan, 2025) and small peptides (Xiao et al., 2025) in influencing stress resilience will help us devise strategies to utilize genomics tools to improve crop plants and crop diversification (Wang et al., 2025).

Here, Krishnan et al. demonstrated the applicability of genomics-assisted tools in a wide-hybridization program involving a heat-tolerant diploid wild wheat *Aegilops speltoides* accession and a *Triticum durum* accession to derive a backcross introgression line (BIL) population. Using this population for marker-assisted selection (MAS), the research group identified 30 QTLs for heat stress tolerance using molecular breeding and genotyping-by-sequencing (GBS) approaches involving both SSR and SNP markers to map the QTLs. Linkage disequilibrium (LD) decay values calculated using Tassel v5.0 helped the authors target 21 candidate genes associated with heat stress tolerance. The most prominent targets based on functional annotation were: cytochrome P450, ABC transporters, E3 ubiquitin-protein ligase, Alcohol dehydrogenase, the F-box family, and

the MYB family. Similarly, in grapevine berries, [Heinekamp et al.](#) used a mapping population (Calardis Musque x Villard Blanc) and important cultivars such as 'Calardis Blanc' (a sunburn-resilient cultivar), totaling a population of 150 for phenotypic evaluation for sunburn resilience in addition to fungal resistance in grapevine cultivars. Using the composite interval mapping (CIM) approach with five years of phenotypic data, along with a genetic map of grapevines, they successfully identified two QTLs that explain approximately 40% of the phenotypic variance and were found to be on chromosomes 10 and 11. With a greater number of heat waves in recent years, climate-change-adapted cultivars that are resistant to fungus and are sunburn resilient are required for sustained viticulture to ensure high yields and wine quality.

Using a population genomics approach, [Zhang et al.](#) demonstrated the influence of low temperature in exerting selection pressure on various plant traits during the adaptation of a *Kandelia obovata* population. To accomplish this, the team introduced a population of *K. obovata* from Zhangzhou (ZZ) to two different locations, Quanzhou (QZ, 2003) and Wenzhou (WZ, 2005). Morphological differences were observed at the two sites. To understand the underlying genetics of this variation, the researchers used a whole-genome resequencing approach and identified the SNPs that varied between the original habitat and the introduced habitat. The positive selection for genes associated with the SNPs (from the northern province WZ) was analyzed to reveal candidate genes linked especially to cold tolerance and glutathione metabolism traits. Analysis of their promoter sequences (extracted up to 2 kbp) showed enrichment of elements primarily linked to stress tolerance, such as stress-, low temperature-, wound-, and abscisic acid (ABA)-responsive elements.

To achieve drought stress resilience in tropical maize and lettuce, [de Pontes et al.](#) and [Medina-Lozano et al.](#) utilized different genomic tools to identify candidate genes in these crops. Using 360 maize inbred lines, SNP-array genotype data from an Affymetrix platform and GBS approaches, the researchers identified drought-associated SNPs and in turn the underlying candidate genes through genome-wide association studies (GWAS). The candidate genes were found to be associated with key pathways such as ethylene biosynthesis, jasmonic acid biosynthesis, gibberellin biosynthesis, and ABA biosynthesis, along with specific protein families such as the TPR, PPR, PR, and MYB families. They were also found to be associated with genes such as shoot gravitropism 5, and circadian clock genes. These findings will contribute to improving maize cultivars tailored with drought resilience for sustained yield improvement. For lettuce, the authors used an RNA-seq approach to study differential expression between the lettuce cultivar 'Romared' and the wild lettuce relative *Lactuca homblei*. The latter is known to significantly overexpress the anthocyanins during drought. Through this study, the authors identified 36 genes, with approximately 50% of them linked to the phenylpropanoid-flavonoid pathway. The other genes were annotated as being associated with stomatal closure, phospholipases, and transcription factors such as MYB, NAC56, PRA1, HSC70, and ZAT1. This provides insight into the regulation of the drought-

mediated anthocyanin pathway for use in imparting drought resilience from the wild species to cultivated lettuce.

In order to promote the molecular breeding and marker-assisted development of novel cultivars of the ornamental plant *Cymbidium ensifolium*, [Shen et al.](#) utilized the double-digest restriction site-assisted DNA sequencing (ddRAD-seq) technique to sequence 50 commercially available cultivars and identified approximately 1.2 million high-quality SNPs. From these SNPs, competitive allele-specific PCR (KASP) primers were designed and used to screen the cultivars and found that 11 of the final 28 KASP markers are sufficient to distinguish the 83 cultivars tested.

In an effort to generate a single circular mitochondrial genome for the decaploid species *Camellia hainanica* (octaploids are also available), [Zhang et al.](#) used Illumina short-read and Nanopore long-read sequencing technologies to generate raw data. After initial filters of the raw data using fastp v0.20.0 and filtlong v0.2.1, the data were mapped against plant mitochondrial core genes using Minimap2 to extract the mitochondrial sequences for assembly using SPAdes v3.15.4, yielding a single circular mitochondrial genome of 902,617 bp in length. The mitochondrial genome was annotated for genes (protein-coding and non-coding) and SSR markers.

The value of conserving germplasm is realized when it is practically utilized in a breeding program to develop cultivars introgressed with key desirable traits from germplasm that would otherwise be difficult to introduce. [He et al.](#) evaluated 361 soybean germplasm accessions, comprising six maturity groups, for variability in 100-seed weight (100SW) and seed oil content (SOC). Using a restricted two-stage multi-locus genome-wide association study (RTM-GWAS) approach, LD blocks comprising 230 and 299 alleles, for 100SW and SOC, respectively, were identified. Gene annotation studies revealed 87 and 132 candidate genes for the 100SW and SOC traits, respectively. Promising 100SW genes included vacuolar proton ATPase A3 and clathrin adaptor complexes. The most promising gene for SOC was identified to be a HAD superfamily phosphatase gene. Genomic selection models using data from the 361 soybean germplasm helped predict the recombination potential of the two studied traits, 100SW (up to 30.43 g) and SOC (up to 27.73%). A model based on priority traits can be chosen for implementation in a breeding program to improve oil yield. [Zhernova et al.](#) reviewed the genetic markers available for the improvement of flax (*Linum usitatissimum*), which is widely cultivated for its oil and fiber. Their review underscores the markers identified for various traits of interest including biotic and abiotic stress tolerance. This compiled information will be ready-to-use for marker-assisted flax improvement.

In conclusion, the enormous growth in data generation and computational power that drives accelerated crop improvement has transformed genomics applications for innovations. Notable advances include the move from single-reference to pan-genome references ([Ruperao et al., 2025](#)), understanding the synergy of the microbiome in crop improvement particularly in terms of yield and stress resilience ([Xu et al., 2025](#); [Ge and Wang, 2025](#)), identifying susceptibility genes ([Baruah et al., 2025](#)), and *de novo* domestication and rational redomestication ([Wang et al., 2025](#)). These advances

will guide research in the coming decades as it addresses the challenges of sustainable agriculture.

Author contributions

PR: Writing – original draft, Writing – review & editing. RH: Writing – review & editing. AG: Writing – review & editing.

Conflict of interest

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

References

- 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, 109–118. doi: 10.1038/s41586-019-1679-0
- Baruah, S., Hutin, M., and Bogdanove, A. J. (2025). Susceptibility genes in bacterial diseases of plants. *Annu. Rev. Phytopathol.* 63. doi: 10.1146/annurev-phyto-121423-104525
- 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, 412–431. doi: 10.1016/j.tibtech.2021.08.009
- Campbell, Q., Bedford, J. A., Yu, Y., Halpin-McCormick, A., Castaneda-Alvarez, N., Runck, B., et al. (2025). Agricultural landscape genomics to increase crop resilience. *Plant Commun.* 6, 101260. doi: 10.1016/j.xplc.2025.101260
- Ge, A.-H., and Wang, E. (2025). Exploring the plant microbiome: a pathway to climate-smart crops. *Cell* 188, 1469–1485. doi: 10.1016/j.cell.2025.01.035
- Gupta, S. (2025). Advances in understanding the functional diversity of noncoding RNA in plants: insights from recent studies. *Ecol. Genet. Genomics* 35, 100360. doi: 10.1016/j.egg.2025.100360
- Hunter, M. C., Smith, R. G., Schipanski, M. E., Atwood, L. W., and Mortensen, D. A. (2017). Agriculture in 2050: recalibrating targets for sustainable intensification. *BioScience* 67, 386–391. doi: 10.1093/biosci/bix010
- Khan, N. (2025). Exploring plant resilience through secondary metabolite profiling: advances in stress response and crop improvement. *Plant Cell Environ.* 48, 4823–4837. doi: 10.1111/pce.15473
- Lohani, N., Singh, M. B., and Bhalla, P. L. (2025). Deciphering the vulnerability of pollen to heat stress for securing crop yields in a warming climate. *Plant Cell Environ.* 48, 2549–2580. doi: 10.1111/pce.15315
- Miryeganeh, M. (2025). Epigenetic mechanisms driving adaptation in tropical and subtropical plants: insights and future directions. *Plant Cell Environ.* 48, 3487–3499. doi: 10.1111/pce.15370
- Ruperao, P., Rangan, P., Shah, T., Sharma, V., Rathore, A., Mayes, S., et al. (2025). Developing pangenomes for large and complex plant genomes and their representation formats. *J. Advanced Res.* doi: 10.1016/j.jare.2025.01.052
- Shrestha, A., Stephens, C. J., and Angeles-Shim, R. B. (2025). Taming wild genomes: perspectives on mechanisms governing differential introgression in exotic rice germplasm and their applications in breeding. *Front. Plant Sci.* 16. doi: 10.3389/fpls.2025.1489244
- Tossolini, I., Mencia, R., Arce, A. L., and Manavella, P. A. (2025). The genome awakens: transposon-mediated gene regulation. *Trends Plant Sci.* doi: 10.1016/j.tplants.2025.02.005
- van Dijk, M., Morley, T., Rau, M. L., and Saghai, Y. (2021). A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* 2, 494–501. doi: 10.1038/s43016-021-00322-9
- Wang, N., Li, H., and Huang, S. (2025). Rational redomestication for future Agriculture. *Annu. Rev. Plant Biol.* 76, 637–662. doi: 10.1146/annurev-arplant-083123-064726
- Xiao, F., Zhou, H., and Lin, H. (2025). Decoding small peptides: regulators of plant growth and stress resilience. *J. Integr. Plant Biol.* 67, 596–631. doi: 10.1111/jipb.13873
- Xu, X., Dinesen, C., Pioppi, A., Kovacs, A. T., and Lozano-Andrade, C. N. (2025). Composing a microbial symphony: synthetic communities for plant growth. *Trends Microbiol.* doi: 10.1016/j.tim.2025.01.006

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Vandna Rai,
Indian Council of Agricultural Research, India
Long Yan,
Hebei Academy of Agriculture and Forestry
Sciences (HAAFS), China

*CORRESPONDENCE

Junyi Gai
✉ sri@njau.edu.cn

RECEIVED 21 August 2024

ACCEPTED 17 October 2024

PUBLISHED 14 November 2024

CITATION

He J, Fu L, Hao X, Wu Y, Wang M, Zhang Q,
Feng W, Fu M, Wang Y, Ren H, Du W,
Wang W and Gai J (2024) Identification of
QTL-allele systems of seed size and oil
content for simultaneous genomic
improvement in Northeast China soybeans.
Front. Plant Sci. 15:1483995.
doi: 10.3389/fpls.2024.1483995

COPYRIGHT

© 2024 He, Fu, Hao, Wu, Wang, Zhang, Feng,
Fu, Wang, Ren, Du, Wang and Gai. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Identification of QTL-allele systems of seed size and oil content for simultaneous genomic improvement in Northeast China soybeans

Jianbo He¹, Lianshun Fu², Xiaoshuai Hao¹, Yicun Wu¹,
Mengfan Wang¹, Qi Zhang¹, Weidan Feng¹, Mengmeng Fu¹,
Yanping Wang³, Haixiang Ren³, Weiguang Du³,
Wubin Wang¹ and Junyi Gai^{1*}

¹Soybean Research Institute & MARA National Center for Soybean Improvement & MARA Key Laboratory of Biology and Genetic Improvement of Soybean & State Key Laboratory of Crop Genetics and Germplasm Enhancement and Utilization & State Innovation Platform for Integrated Production and Education in Soybean Bio-Breeding & Zhongshan Biological Breeding Laboratory & Jiangsu Collaborative Innovation Center for Modern Crop Production, Nanjing Agricultural University, Nanjing, China, ²Soybean Research Institute, Tieling Academy of Agricultural Sciences, Tieling, China, ³Mudanjiang Research and Development Center for Soybean & Mudanjiang Experiment Station of the National Center for Soybean Improvement, Mudanjiang Branch of Heilongjiang Academy of Agricultural Sciences, Mudanjiang, China

Northeast China (NEC) is the major production area for soybeans in China, whereas its soybean germplasm has played key roles in world soybean production, especially in the Americas. For plant breeding, genomic selection involves two stages, cross design and progeny selection, with the former determining the latter's potential. In NEC, one of the major breeding purposes is for 100-seed weight (100SW) and seed oil content (SOC). A diverse sample with 361 NEC soybean germplasm accessions was evaluated for their 100SW and SOC in Tieling, Liaoning, China. Both traits exhibited significant phenotypic, genotypic, and G × E variation, with a trait heritability of 82.38% and 86.26%, respectively. A restricted two-stage multi-locus genome-wide association study (RTM-GWAS) with 15,501 SNPLDB (SNP linkage disequilibrium block) markers identified 80 and 92 QTLs with 230 and 299 alleles for 100SW and SOC, respectively. Corresponding to some increase of the two traits, almost all the alleles in the early maturity groups (MG 0 + 00 + 000) were inherited from the late MGs (MG I + II + III), indicating that genetic recombination was the major motivator in addition to a few allele emergence and some allele exclusion fluctuations among early MGs. Using the 95th percentile as indicator, the prediction of

recombination potentials showed that 30.43 g 100SW and 27.73% SOC might be achieved, respectively. Three strategies of simultaneous genomic improvement of both traits in designing optimal crosses, namely, 100SW-first, SOC-first, and 100SW-SOC-balance, were proved to be efficient. Thus, the optimal cross design could be extended to multiple traits based on a relatively thorough identification of the QTL-alleles using RTM-GWAS.

KEYWORDS

Northeast China soybean germplasm population (NECSGP), 100-seed weight (100SW), seed oil content (SOC), RTM-GWAS, recombination potential prediction, simultaneous genomic cross design for traits, evolutionary dynamics

1 Introduction

Northeast China (NEC) is the major production area for soybeans, with its yield and acreage accounting for approximately 50% in China. Liu et al. (2020) showed that soybeans in North and South America were genetically clustered in the same group with those from NEC, and the germplasm from NEC is the primary source for soybeans in the world's largest production regions. Therefore, exploring the genetic basis of NEC soybean germplasm is essential for global soybean production (Fu et al., 2020a). During the long period of artificial selection and improvement in soybean, the seed size (expressed as 100-seed weight or 100SW), seed oil content (SOC), and seed protein content (SPC) increased, but there was a negative correlation between oil and protein contents (Guo et al., 2022; Li et al., 2022; Ray et al., 2022). Compared with soybeans in central and southern China, soybeans in NEC have relatively smaller 100SW but higher SOC; thus, one of the major breeding purposes is for increased 100SW and SOC.

Marker-assisted selection (MAS) has been proven as an effective method for precise plant breeding (Collard and Mackill, 2008). Through identifying genomic markers related to breeding targets, MAS improves breeding efficiency by applying directly genotypic selection in addition to phenotypic selection. Furthermore, the concept of “breeding by design” was proposed based on quantitative trait locus (QTL) and gene mapping, aiming to improve breeding efficiency by designing optimal genotypes and parental combinations (Peleman and van der Voort, 2003). However, breeding by design is usually applicable for the selection of a handful of major genes. Meanwhile, genomic selection (GS) was also proposed to improve breeding efficiency by predicting breeding values in offspring population using whole-genome markers without the need for QTL/gene mapping (Meuwissen et al., 2001). GS firstly establishes a statistical model between phenotype and genome-wide markers in a training or reference population (generally germplasm population) where both phenotype and genotype data are available and then predicts breeding values as a comprehensive evaluation for all target traits in breeding populations using genome-wide markers based on an

established model. In fact, both breeding by design and GS are special cases of MAS in comprehensive selection for multiple traits. Following the “breeding by design” concept, QTL-allele-based GS was also proposed as a potential approach to both optimal crosses and superior progenies based on the whole-genome QTL-allele system, and is in fact a direct genotype selection method (He et al., 2017). Anyway, simultaneous improvement of multiple traits using QTL-allele-based GS in plant breeding needs to be further explored and practiced.

Both 100SW and SOC are quantitatively inherited traits controlled by a number of genes with varying effects. At least 297 and 315 QTLs have been reported at SoyBase (<https://www.soybase.org>) for 100SW and SOC in soybean, respectively. These QTLs were detected mainly from bi-parental populations using the linkage mapping method with many ones located in overlapping regions. However, among these, only 13 and 16 QTLs of 100SW and SOC were recognized as confirmed QTLs, respectively (Fasoula et al., 2004; Nichols et al., 2006; Pathan et al., 2013). There were also a handful of genes conferring the two traits, which have been cloned. For example, a wild soybean allele simultaneously conferring 100SW, seed protein, and oil content was mapped to a 329-kb region on chromosome 15, with *Glyma.15g049200* as their common candidate gene (Yang et al., 2019). The causative gene *SW16.1* underlying a 100SW QTL *qSW16.1* was identified and encoded as a nucleus-localized LIM domain-containing protein (Chen et al., 2023). However, the underlying genes for most of the detected QTLs still need to be further identified; in particular, the complete gene systems of 100SW and SOC QTLs need to be explored for a thorough genetic design.

The linkage mapping for QTL detection involves only two parental lines, such as the recombinant inbred line (RIL) population, in which the genetic variation is limited between the two parents. The genome-wide association study (GWAS) for the natural/germplasm population provides a broad background to genome-wide QTL identification (Hong et al., 2023; Khatun et al., 2022; Ren et al., 2022; Ullah Zaid et al., 2018) and was widely used in soybean for QTL identification (Li et al., 2017; Zhang et al., 2018;

Su et al., 2023). The GWAS on over 12,000 soybean accessions identified 18 and 19 QTLs for SOC and SPC, respectively (Bandillo et al., 2015). Based on 809 soybean accessions worldwide, GWAS identified 245 QTLs of 84 agronomic traits (Fang et al., 2017). Hwang et al. (2014) identified 25 and 40 SOC and SPC loci in 298 soybean germplasm accessions, respectively. The GWAS also provided an efficient method for mining the underlying genes conferring quantitative trait variation. A sucrose efflux transporter gene, *GmSWEET39*, controlling SOC was identified using GWAS (Miao et al., 2020). A pair of *SWEET* homologs, *GmSWEET10a* and *GmSWEET10b*, were identified simultaneously conferring 100SW, SOC, and SPC during soybean domestication (Wang S. et al., 2020).

Generally, only a handful of major QTLs were detected in individual GWAS, whereas their multiple alleles in germplasm populations were neglected. To improve the GWAS efficiency, the restricted two-stage multi-locus genome-wide association analysis (RTM-GWAS) was proposed for a relatively thorough detection of whole-genome QTLs with their multiple alleles (He et al., 2017). Two innovative techniques in RTM-GWAS were taken. One method uses SNPLDB (SNP linkage disequilibrium block) markers with multiple haplotypes to meet the requirements of multiple alleles in the natural population. The other method controls the total contribution within heritability value based on two-stage GWAS with the first stage under a single-locus model for marker pre-selection and the second stage of stepwise regression under a multi-locus model with forward addition and backward elimination. RTM-GWAS was used in QTL-allele detection for 100SW and SOC in 366 soybean landraces (Zhang et al., 2015, 2018). Firstly, 116,769 single-nucleotide polymorphism (SNP) markers were used to form 29,121 SNPLDBs. Then, 55 and 50 QTLs with 263 and 136 alleles were detected based on SNPLDBs for 100SW and SOC, respectively. The detected QTL-allele matrix was used in studies for evolutionary motivators and breeding potentials of the respective trait.

Conventional crossbreeding has been the major procedure for genetic improvement in soybean. It consists of two major steps: the first is to design optimal crosses based on germplasm or breeding materials, and the second is to select the best progenies in segregating generations. Optimal cross design determines the potential of progeny selection and is the key to breakthrough breeding. An approach of GS based on the QTL-allele matrix was proposed for optimal cross design (He et al., 2017; He and Gai, 2020) and has been applied to soybean breeding (Khan et al., 2018; Wang W. et al., 2020; Fahim et al., 2021; Wang et al., 2021). For example, 1,803 optimal crosses were predicted for high SPC in the NEC soybean germplasm population (NECSGP), and the maximum of predicted SPC was 50.00% with a transgressive potential of 3.93% improvement (Feng et al., 2022).

However, previous studies on cross design based on GWAS-identified QTL-allele results focused on genetic improvement of a single trait. Since breeding programs involve multiple traits, it is essentially appropriate to consider multiple traits simultaneously in designing optimal crosses. In this study, the representative sample of the NECSGP (Feng et al., 2022) was studied to identify the 100SW and SOC QTL-allele systems expressed in NEC using the RTM-GWAS procedure, to characterize the genetic systems and

motivators of the two traits in the evolutionary process from the southern part to the northern part of NEC, and to predict the recombination potential or to design optimal crosses for the simultaneous improvement of the two traits in the NECSGP using 100SW-first, SOC-first, and 100SW-SOC-balance strategies, respectively. This optimal cross design method could be extended to multiple traits based on a relatively thorough identification of the QTLs and their alleles of the traits.

2 Materials and methods

2.1 Plant materials and field experiments

As described previously (Feng et al., 2022), a total of 361 soybean accessions from the NEC were used in this study. These accessions involve six soybean maturity groups, namely, III, II, I, 0, 00, and 000 (Fu et al., 2020a). Accessions in MG III matured later, whereas accessions in MG 000 matured earlier. Field experiments were performed at Tieling, Liaoning in 2013–2014. All accessions were grouped into six blocks according to maturity group and were planted using the “blocks in replication” design with four replications. Normal field management including weed control and fertilization was used. The matured seeds were harvested and dried under 35–40°C. The 100SW (g) was measured on 100 randomly selected seeds, and the SOC was determined by the near-infrared grain analyzer Infratec 1241 (FOSS, Hilleroed, Denmark).

2.2 Statistical analysis

Statistical analysis was performed using the SAS Studio software through SAS OnDemand for Academics (<https://welcome.oda.sas.com/>). Joint analysis of variance (ANOVA) was conducted using the PROC GLM procedure, while the variance components were calculated using PROC VARCOMP in which genotype, environment, replication within environment, and genotype $\times$ environment were considered as random effects. The heritability (h^2) was estimated as $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma^2 / r)$ for the single environments and $h^2 = \sigma_g^2 / (\sigma_g^2 + \sigma_{gt}^2 / t + \sigma^2 / tr)$ for the combined data over multiple environments, where σ_g^2 , σ_{gt}^2 and σ^2 are variance of the genotype, genotype $\times$ environment, and random error; t is the number of environments; and r is the number of replications (Hanson et al., 1956).

2.3 Genotyping and haplotype block marker construction

The genotype data of the 361 accessions were obtained from Fu et al. (2020a). Whole-genome sequencing of the accessions were carried out using RAD-seq (restriction-site associated DNA sequencing) at BGI Tech, Shenzhen, China. The genomic DNA was isolated from the young leaves of soybean seedlings according to the conventional CTAB method (Murray and Thompson, 1980).

Paired-end sequencing was conducted on an Illumina HiSeq2000 platform through the multiplexed shotgun genotyping method (Andolfatto et al., 2011). All sequence reads were aligned against the genome of Williams 82 using the SOAP2 software (Li et al., 2009; Schmutz et al., 2010). RealSFS (Yi et al., 2010) was utilized to detect SNP loci, which were filtered with a maximum missing and heterozygous allele call rate of $\leq 20\%$ and a minimum minor allele frequency (MAF) of $\geq 1\%$. The fastPHASE software (Scheet and Stephens, 2006) was used for genotyping the SNP imputation resulting in 82,966 high-quality SNPs. The SNPs after quality control were grouped into 15,501 SNPLDBs based on the LD threshold of $D' \geq 0.7$ according to He et al. (2017), while haplotypes were treated as alleles of a QTL.

2.4 Detection of the QTL-allele system and the establishment of matrix/sub-matrices

The RTM-GWAS software (<https://gitee.com/njau-sri/rtm-gwas>) was used to detect the QTL-allele system of 100SW and SOC in the NECSGP. The genetic similarity between accessions was estimated based on genome-wide SNPLDBs, and the top 10 eigenvectors of the genetic similarity coefficient matrix were used as the covariates to correct the population structure bias. A threshold of $p = 0.05$ was used at the first stage of RTM-GWAS for candidate marker preselection, and a significance level of $p = 0.05$ was used for QTL detection through stepwise regression at the second stage of RTM-GWAS. The detected QTLs (associated SNPLDBs) with their allele effects for each accession were used to establish the QTL-allele matrix, which was further separated into maturity group sub-matrices for further analysis on evolutionary motivators.

2.5 Candidate gene annotation

The candidate genes for 100SW and SOC from the detected QTLs were annotated through the following steps: (1) the genes of soybean genome Wm82.a1.v1.1 within the genomic interval of a detected QTL (with a 50-kb flanking expansion) were retrieved from SoyBase (<https://www.soybase.org>); (2) the independence of SNP(s) between an identified SNPLDB and gene(s) within the genomic interval was statistically tested using the chi-square criterion at a significance level of 0.05; and (3) the Gene Ontology (GO) annotations of significantly correlated genes were retrieved from SoyBase (<https://www.soybase.org>).

2.6 Recombination potential prediction and optimal cross design

All possible 64,980 crosses ($361 \times 360/2$) were generated *in silico*. For each cross, the genotype data of 2,000 homozygous progenies were simulated through continuous selfing starting from F_2 generation under the linkage model, where the number of crossovers on each chromosome was simulated randomly

according to the Poisson distribution with chromosome length as a parameter. The predicted genotypic values of each progeny were calculated as the sum of all allele effects plus the population mean (He et al., 2017). Different percentiles of the progeny population were calculated and used as the predicted recombination potential. The *cross* program (<https://gitee.com/njau-sri/cross>) was used for simulation. In addition, the crosses were also grouped according to maturity groups, then the recombination potential within and among maturity groups was also analyzed.

Three strategies were proposed in this study for two-trait optimal cross design of soybean with both high 100SW and SOC. In the first strategy (designated 100SW-first), the top 100 crosses with the highest recombination potential for 100SW were selected firstly, and then the top 10 crosses with the highest recombination potential for SOC were selected from the 100 crosses. In the second strategy (designated SOC-first), the top 100 crosses with the highest recombination potential for SOC were selected firstly, and then the top 10 crosses with the highest recombination potential for 100SW were selected from the 100 crosses. The third strategy (designated 100SW-SOC-balance) involved balanced selection for 100SW and SOC; firstly, all possible crosses were arranged in descending order according to the recombination potential for 100SW and SOC, respectively, and then the top 10 most common crosses were selected from the top 2,000 100SW and SOC crosses.

3 Results

3.1 Variation of 100SW and SOC in the NECSGP

The joint ANOVA over 2 years indicated that the 100SW and SOC varied both significantly among accessions (genotypes), as well as their genotype $\times$ environment interactions (GEI, Supplementary Table 1) in the NECSGP. However, the estimated GEI variances were relatively small compared to the genotypic variance, indicating that the interaction between genotype and environment (year) was weak for the two traits in the NECSGP. Both 100SW and SOC exhibited higher heritability values over two environments, 82.46% and 86.34, respectively. The 100SW in the NECSGP ranged from 9.00 to 27.20 g across 2 years, with an average of 18.37 g (Table 1; Figure 1A), which was not as wide as that in the Chinese soybean landrace population (CSLP, ranging from 4.59 to 40.35 g in Zhang et al., 2015). The SOC in the NECSGP ranged from 18.80% to 24.85% across 2 years, with an average of 22.45% (Table 1; Figure 1E), which was similar to the CSLP (ranging from 14.95% to 26.42%) in Zhang et al. (2018). The genetic coefficient of variation (GCV) of SOC varied relatively less (4.12%) than that of 100SW (11.77%) (Table 1). The above results indicate that the variability of 100SW and SOC in the NECSGP is not better than those of CSLP; the improvement of the two traits needs to explore their recombination potential.

There were some variations of 100SW and SOC among different maturity groups in the NECSGP, but the major difference was within maturity groups. Overall, the 100SW and SOC in the late maturity groups (MG I+II, except for III for the former, and MG II

TABLE 1 The distribution and descriptive statistics of 100-seed weight and seed oil content in the NECSGP.

Trait	Year	Midpoint													N	Mean	Range	GCV (%)	h ² (%)
100SW (g)		9.0	10.5	12.0	13.5	15.0	16.5	18.0	19.5	21.0	22.5	24.0	25.5	27.0					
	2013	1	1	1	2	8	57	90	83	66	30	14	6	2	361	19.31	9.18–27.20	12.50	87.08
	2014	1	1	2	22	45	107	86	55	27	10	5			361	17.44	9.00–24.63	13.06	79.67
	Mean	1	1	0	5	22	87	97	84	38	18	7	1		361	18.37	9.09–25.16	11.77	82.38
SOC (%)		19.0	19.5	20.0	20.5	21.0	21.5	22.0	22.5	23.0	23.5	24.0	24.5	25.0					
	2013	1	5	7	21	16	38	60	64	74	46	21	7	1	361	22.39	18.80–24.75	4.79	92.41
	2014		2	1	12	15	37	58	88	70	42	27	4	1	357	22.52	19.55–24.85	3.99	92.25
	Mean	1	0	3	15	16	37	71	78	62	59	13	6		361	22.45	19.21–24.73	4.12	86.26

100SW, 100-seed weight; SOC, seed oil content. N, the number of accessions. GCV, the genetic coefficient of variation defined as genetic standard deviation divided by phenotype mean. *h*², trait heritability. The SOC was evaluated with the near-infrared grain analyzer Infratec 1241 (FOSS, Hilleroed, Denmark), and its value might be inflated at some degree. However, all the evaluation of SOC was kept under the same environment and comparable.

+III, except for I for the latter) were smaller than those in early maturity groups (MG 000 + 00 + 0) (Supplementary Table 2). This suggested that both 100SW and SOC had been genetically improved in the NECSGP when adapting to high latitude with days to maturity shortened.

The top 10 100SW accessions (F67, F326, F406, F147, F58, P004, P085, F36, P188, and F315 in descending order) ranged from 25.16 to 23.05 g (average for 2 years), with their SOC ranging from 22.60% to 20.38%. On the other hand, the top 10 SOC accessions (F82, F386, F135, F79, F32, F155, F53, F305, F306, and F351 in descending order) ranged from 24.73% to 24.09% (average for 2 years), with their 100SW ranging from 21.04% to 15.64%. No accession had both the highest 100SW and the highest SOC in the NECSGP. The correlation analysis showed that 100SW and SOC did not exhibit a significant Pearson correlation coefficient (*r* = 0.008, *p* = 0.8722), suggesting that simultaneous improvement of seed size and oil content might be possible in the NECSGP. To effectively improve both 100SW and SOC in soybean breeding, a thorough detection of the whole-genome QTL-allele constitution for these two traits is an essential requirement.

3.2 Genome-wide detection of 100SW and SOC QTL-allele systems in the NECSGP

As described previously in Feng et al. (2022), a total of 15,501 SNPLDBs were constructed from 82,966 SNPs in a RAD-seq procedure in the NECSGP. In the first stage of RTM-GWAS under the single-locus model, 8,272 and 8,863 out of 15,501 SNPLDBs were preselected for 100SW and SOC, respectively. In the second stage under the multi-locus model, 80 and 92 QTLs were detected for 100SW and SOC, respectively (Figures 1B, C, F, G).

For 100SW, 80 QTLs were identified (Supplementary Table 3; Figure 1D). These QTLs were distributed on 19 chromosomes except for chromosome 16, with seven QTLs located on chromosome 13 being the highest. Among the 80 QTLs, 42 had the main effect only, 8 QTLs had the QEI effect only, and 30 QTLs

had both the main effect and QEI effect. The phenotypic contribution of the 72 QTLs with a significant main effect varied continuously from 0.06% to 6.58% in a total of 54.11% phenotypic variation (Table 2). The 16 large-contribution QTLs (main effect *R*² ≥1%, around the inflection point of the phenotypic contribution curve) explained 34.60% phenotypic variation, and 56 small-contribution QTLs (main effect *R*² <1%) explained 19.51% phenotypic variation. A total of 230 alleles were identified on the 80 QTLs, with the number of alleles per QTL ranging from 2 to 8, and 27 QTLs had at least 3 alleles (Figure 1I).

For SOC, 92 QTLs were identified (Supplementary Table 4; Figure 1H). These QTLs were distributed on all 20 chromosomes, with 10 QTLs on chromosome 9 being the highest. Among the 92 QTLs, 38 had the main effect only, 10 QTLs had only the QEI effect, and 44 QTLs had both the main effect and QEI effect. Thus, 82 QTLs had significant main effects that varied continuously from 0.04% to 4.70% in a total of 70.07% phenotypic variation. The 25 large-contribution QTLs explained 53.10% and 57 small-contribution QTLs explained 16.96% phenotypic variation (Table 2). There were 299 alleles on the 92 QTLs, with the number of alleles per QTL ranging from 2 to 8, and 38 QTLs with 3 or more alleles (Figure 1J).

The allele effect of main effect QTLs ranged from −4.53 to 3.00 g for 100SW and from −1.79 to 1.75% for SOC (Figures 1I, J). The 72 and 82 main effect QTLs and their allele effects for each of the 361 accessions were organized as a QTL-allele matrix, respectively (Figures 2A, B), a compact form of the genetic constitution of the two traits in the NECSGP. In addition, the QTL-allele matrix was further separated into sub-matrices corresponding to the six maturity groups for comparisons and evolutionary changes of 100SW and SOC among maturity groups.

The QTL-allele constitution of the top 10 accessions of the highest 100SW and SOC is shown in Figures 2C, D, respectively. These accessions contained many alleles of negative effects on most QTLs, indicating large improvement potential for both 100SW and SOC.

In the following text, the focus will be on the main effect QTL-allele matrices, while the GEI matrices will be left for future analysis,

TABLE 2 Summary of the QTL-allele system for 100-seed weight and seed oil content in the NECSGP.

QTL-allele	100-seed weight		Seed oil content	
	Main effect	QTL×Year	Main effect	QTL×Year
QTL (R^2 , %)	54.11 (72, 0.06–6.58)	7.34 (38, 0.08–0.53)	70.06 (82,0.04–4.70)	9.08 (54, 0.03–0.75)
LC QTL (R^2 , %)	34.60 (16, 1.05–6.58)		53.10 (25, 1.02–4.70)	
SC QTL (R^2 , %)	19.51 (56, 0.06–0.95)	7.34 (38, 0.08–0.53)	16.96 (57, 0.04–0.97)	9.08 (54, 0.03–0.75)
Unmapped QTL (R^2 , %)	28.35	0.58	16.28	0.27
Heritability (h^2 , %)	82.46	7.92	86.34	9.35
Total allele	232 (–4.53–3.00)	464 (–1.75–1.75)	299 (–1.79–1.75)	598 (–1.11–1.11)
Positive allele	121 (0.00–3.00)	232 (0.00–1.75)	161 (0.00–1.75)	299 (0.00–1.11)
Negative allele	111 (–4.53–0.04)	232 (–1.75–0.00)	138 (–1.79–0.00)	299 (–1.11–0.00)

Main effect: main-effect QTL. QTL×Year: QTL by Year interaction. R^2 : phenotypic variation explained. LC QTL: large-contribution QTL ($R^2 \geq 1\%$). SC QTL: small-contribution QTL ($R^2 < 1\%$). In parentheses of QTL rows, the first number is the number of identified QTLs, followed by a range of single QTL contributions to phenotypic variance. In parentheses of allele rows is the range of single allele effects.

as the environment factor of year varied not definitely but randomly.

3.3 Annotation of the candidate gene system of 100SW and SOC in the NECSGP

Among the detected 100SW QTLs, 87 genes were annotated from 37 QTLs, while no candidate genes were annotated for the remaining 43 QTLs, according to SoyBase (<https://www.soybase.org>). These 87 genes were distributed on 17 chromosomes, excluding Gm01, Gm08, and Gm16. Similarly, 132 genes were annotated from 40 SOC QTLs, while no candidate genes were annotated for the remaining 52 SOC QTLs. These 132 genes were distributed on 16 chromosomes, excluding Gm01, Gm05, Gm08, and Gm16. GO annotations indicated that these genes involved cellular components, molecular functions, and biological processes, especially the latter, including 16 function groups (Figure 1K). This suggested that the candidate gene systems of the two traits involved multiple different groups of functions. The QTL *qSW-11-2* with the highest explained phenotypic contribution to 100SW ($R^2 = 6.58$) had nine associated genes, among which *Glyma11g12120* (vacuolar proton ATPase A3) and *Glyma11g12230* (clathrin adaptor complexes medium subunit family protein) were highly expressed in soybean pod and seed (SoyBase, <http://www.soybase.org>). In the SOC QTL *qSOC-18-2* with the highest explained phenotypic variation ($R^2 = 4.7\%$), three genes were annotated including *Glyma18g03450*, a HAD superfamily phosphatase highly expressed in soybean pod and seed. In addition, based on the gene models for which there was a significant change of gene expression in seed developmental stages (Severin et al., 2010), three genes (*Glyma09g25490*, *Glyma10g05580*, and *Glyma13g36780*) and four genes (*Glyma07g32050*, *Glyma09g32980*, *Glyma11g10600*, and *Glyma19g42380*) were identified from the annotated genes for

100SW and SOC, respectively. In particular, two genes, *Glyma09g32980* and *Glyma11g10600*, were annotated from two large-contribution SOC QTLs (*qSOC-9-8* with $R^2 = 4.07$ and *qSOC-11-1* with $R^2 = 3.87$). These findings provide insights into the genetic architecture underlying 100SW and SOC and highlight specific genes that may be important for these traits.

3.4 Genetic differentiation of QTL-allele from late to early maturity groups

During the evolutionary process from the late (MG I+II+III) to early (MG 0 + 00 + 000) maturity groups, all 225 alleles for 100SW and 292 out of 294 alleles for SOC were retained (Table 3), suggesting that inheritance or migration was the major genetic dynamics for these two traits. Seven positive effect alleles emerged in the early maturity groups for 100SW. Two negative and three positive alleles emerged, and one negative and one positive allele were excluded in the early maturity groups for SOC. Between MG 0 + 00 + 000 and MG I+II+III, only a few allele changes happened for the two traits. However, the QTL-allele structure of both 100SW and SOC exhibited genetic fluctuation among the three early maturity groups, particularly with a large number of excluded alleles. Among the early maturity groups, MG 000 inherited only 156 100SW alleles and 186 SOC alleles from the late maturity groups (MG I+II+III), with 30% (69/225) of the 100SW alleles and 37% (108/294) of the SOC alleles excluded, while newly emerged alleles increased very little. In this case, for the two major genetic dynamics, allele emergence was quite limited whereas allele exclusion fluctuated; therefore, genetic recombination among alleles on different QTLs might play an important role in generating phenotypic variation among maturity groups. This might be the major genetic basis of the predicted transgressive optimal crosses with linkage obstacles deleted.

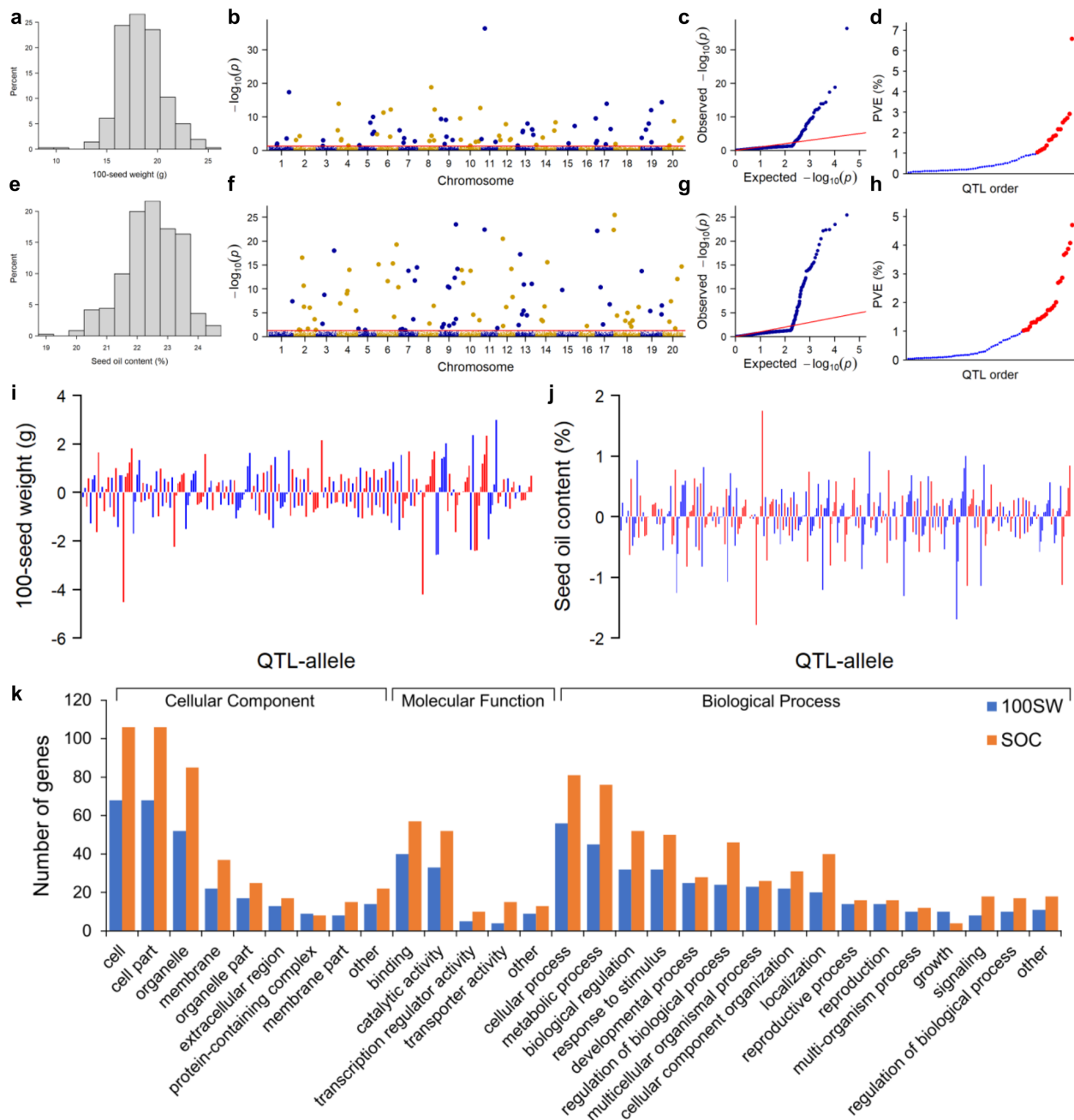


FIGURE 1

QTL-allele detection for 100-seed weight and seed oil content in the NECSGP. (A–D) Genome-wide association study of 100-seed weight: histogram of phenotype data (A), Manhattan plot (B) and quantile–quantile plot (C) of marker p -values, and phenotypic contribution (R^2) of the main-effect QTLs in ascending order (D) with blue color for small-contribution QTL ($R^2 < 1\%$) and red color for large-contribution QTL ($R^2 \geq 1\%$). (E–H) Genome-wide association study of seed oil content: histogram of phenotype data (E), Manhattan plot (F) and quantile–quantile plot (G) of marker p -values, and phenotypic contribution (R^2) of the main-effect QTLs in ascending order (H) with blue color for small-contribution QTL ($R^2 < 1\%$) and red color for large-contribution QTL ($R^2 \geq 1\%$). (I) Allele effects of main-effect QTLs for 100SW. (J) Allele effects of main-effect QTLs for SOC. (K) Candidate gene annotation for 100SW and SOC QTLs.

3.5 Optimal cross design for the improvement of individual and simultaneous improvement of 100SW and SOC in the NECSGP

The genetic recombination potential for 100SW and SOC in the NECSGP was predicted based on their respective QTL-allele matrix.

As the breeding target is high 100SW or SOC, the 95th percentile of predicted 100SW or SOC in the progeny population was used as an indicator for the recombination potential of a parental cross in the NECSGP. As shown in Table 4 and Figures 2E, F, the transgressive potential of the progenies for all possible 64,980 crosses among the 361 accessions was predicted for the respective traits in the NECSGP. For 100SW, the average recombination potential was

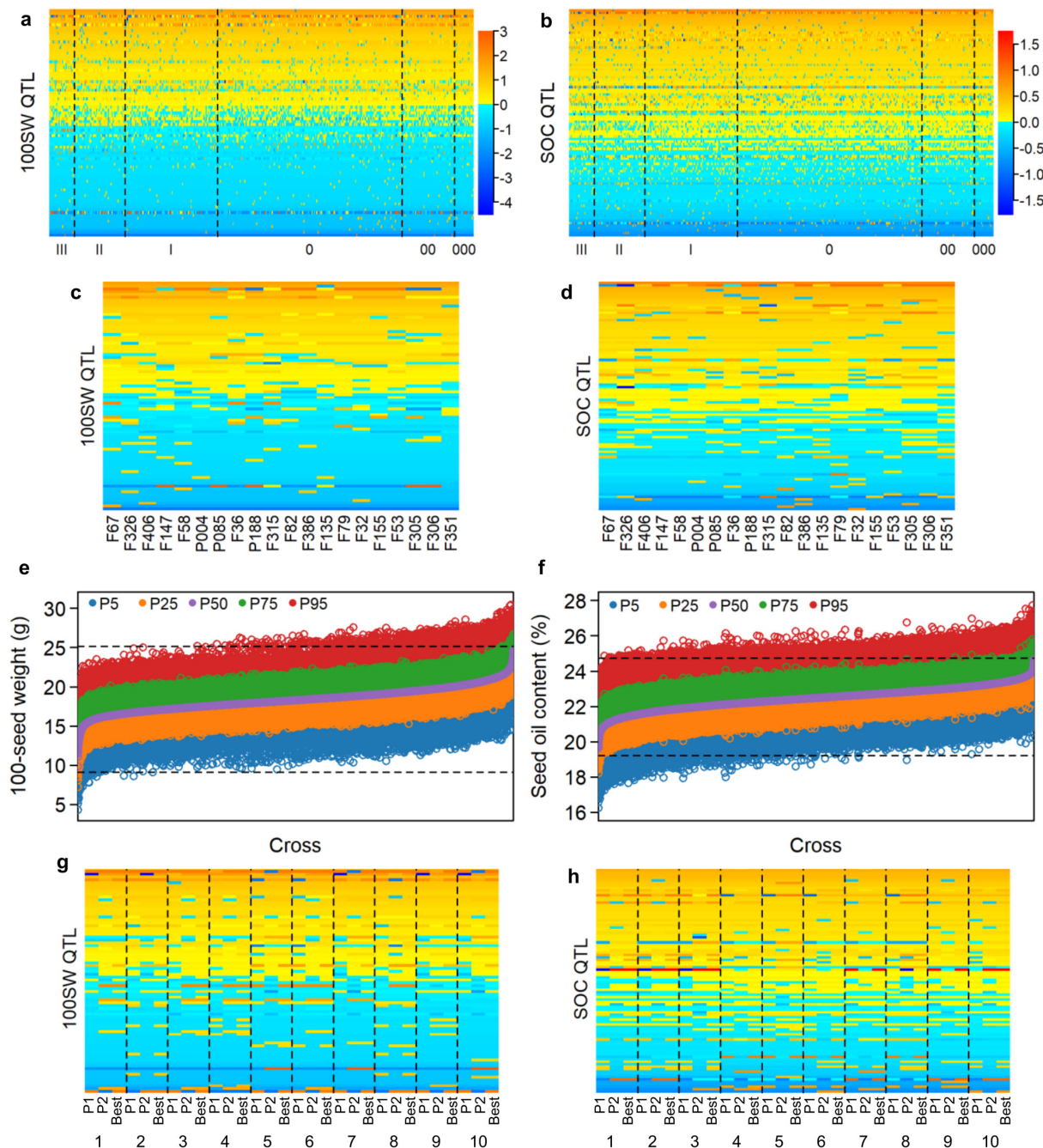


FIGURE 2

QTL-allele matrix and recombination potential of 100-seed weight and seed oil content in the NECSGP. (A) Graphical representation of the QTL-allele matrix of 100SW. The horizontal axis represents accessions, while the vertical axis represents QTL arranged in a rising order of their positive allele frequency. Each row represents the allele distribution among accessions for a QTL, while each column indicates the allele constitution of an accession over all QTLs. Allele effects are expressed in color cells with warm colors indicating positive effects and cool colors indicating negative effects, and the color depth indicates effect size. (B) Graphical representation of the QTL-allele matrix of SOC. (C) The QTL-allele matrix of 100SW for the 20 best accessions. (D) The QTL-allele matrix of SOC for the 20 best accessions. (E) Distribution of predicted 100SW of progenies in all possible crosses, with the maximum and minimum (upper and lower horizontal dotted lines). (F) Distribution of predicted SOC of progenies in all possible crosses, with the maximum and minimum (upper and lower horizontal dotted lines). (G) Graphical representation of the QTL-allele matrix of 100SW for the top 10 optimal crosses (1–10 in Supplementary Table 5) each with P1, P2, and the best progeny. (H) Graphical representation of the QTL-allele matrix of SOC for the top 10 optimal crosses (1–10 in Supplementary Table 6) each with P1, P2, and the best progeny.

21.62 g, reaching a maximum of 30.42 g. This surpassed the observed 100SW in the NECSGP, which had a mean of 18.37 g and a maximum of 25.16 g, leading to a maximum improvement. Similarly, the average recombination potential for SOC was 24.42%,

reaching a maximum of 27.73%, surpassing the observed average (22.45%) and maximum (24.73%) in the NECSGP.

Among the top 100 optimal crosses out of the 64,980 ones for 100SW, 13 crosses had both parents from the top 10 100SW

TABLE 3 The QTL-allele changes of 100 seed weight and seed oil content among maturity groups.

Trait	Maturity group	Total		Inherent		Emerged		Excluded	
	Contrast	Allele	QTL	Allele	QTL	Allele	QTL	Allele	QTL
100SW	I+II+III	225 (111,114)	80						
	0 vs. I+II+III	226 (109,117)	80	219 (109,110)	80	7 (0,7)	7	6 (2,4)	6
	00 vs. I+II+III	196 (92,104)	80	190 (92,98)	80	6 (0,6)	6	35 (19,16)	27
	000 vs. I+II+III	156 (72,84)	80	156 (72,84)	80	0 (0,0)	0	69 (39,30)	52
	0 + 00 + 000 vs. I+II+III	232 (111,121)	80	225 (111,114)	80	7 (0,7)	7	0 (0,0)	0
SOC	I+II+III	294 (136,158)	92						
	0 vs. I+II+III	293 (133,160)	92	288 (131,157)	92	5 (2,3)	4	6 (5,1)	5
	00 vs. I+II+III	239 (97,142)	92	235 (96,139)	92	4 (1,3)	4	59 (40,19)	51
	000 vs. I+II+III	188 (81,107)	92	186 (81,105)	92	2 (0,2)	2	108 (55,53)	74
	0 + 00 + 000 vs. I+II+III	297 (137,160)	92	292 (135,157)	92	5 (2,3)	4	2 (1,1)	2

100SW, 100-seed weight; SOC, seed oil content. The number outside parentheses is the total of alleles; the numbers in parentheses are the number of negative effect alleles and positive effect alleles, respectively. Inherent allele means alleles passed from the compared maturity group(s); Emerged allele means the alleles new to the compared maturity group(s); Excluded allele means the alleles excluded in the maturity group(s).

accessions, and 52 crosses had only one parent from the top 10 100SW accessions (Supplementary Table 5; Figure 2G). There were 35 crosses that did not include any of the top 10 100SW accession. Similarly, among the top 100 optimal crosses for SOC, there were 14 crosses with both parents from the top 10 SOC accessions, and 48 crosses with only one parent from the top 10 SOC accessions (Supplementary Table 6; Figure 2H). Additionally, 38 crosses did not involve any of the top 10 SOC accessions. Even some were involved with the top ones, but not ranking in the front position. This indicated that the top 10 accessions might play an important role in breeding for 100SW and SOC, but a breakthrough improvement may not be readily achieved through crossing the best accessions only. On the other hand, the best predicted cross for SOC had its parent without any of the top 10 SOC accessions. This indicated that complementary potential lay in specific pairs of the accessions, not necessarily in only the best accessions of the

TABLE 4 Predicted 95th percentile of 100SW and SOC of all possible crosses in/between maturity groups in the NECSGP.

Maturity group	N	100-seed weight (g)			Seed oil content (%)		
		Min.	Max.	Mean	Min.	Max.	Mean
Entire	64,980	15.76	30.43	21.62	21.32	27.73	24.20
I+II+III	10,153	16.03	29.54	21.21	21.32	27.14	23.94
0 + 00 + 000	23,653	16.72	28.57	21.84	21.81	27.12	24.35
0	12,246	16.48	29.52	21.68	22.07	27.12	24.37
00	990	17.48	27.74	22.38	22.16	26.44	24.39
000	120	18.89	26.31	21.90	22.44	25.07	23.93
0 + 00 + 000 vs. I+II+III	31,174	15.76	30.43	21.59	21.54	27.73	24.17
0 vs. I+II+III	22,451	15.98	30.43	21.49	21.54	27.42	24.18
00 vs. I+II+III	6,435	15.76	29.99	21.89	22.16	27.73	24.21
000 vs. I+II+III	2,288	17.52	29.27	21.67	21.78	26.73	23.98
0 vs.00	7,065	16.58	28.87	22.04	21.99	27.10	24.40
0 vs. 000	2,512	17.66	28.62	21.79	21.81	26.41	24.16
00 vs. 000	720	18.08	27.93	22.11	22.12	26.42	24.15

N: The number of crosses. SD, standard deviation. The predicted phenotypic value for each cross was defined as the 95th percentile of the predicted progeny. 0 + 00 + 000 vs. I+II+III means crosses between 0 + 00 + 000 and I+II+III, and the same is true for the others.

NECSGP. Therefore, evaluation of each accession's QTL-allele constitution for finding the best complementary potential pairs is necessary in a germplasm/breeding population. In other words, for 100SW and SOC in the NECSGP, "specific combining ability" or "specific recombination potential" is more important.

Although there were significant differences in 100SW and SOC among different maturity groups, the predicted recombination potentials within MG I+II+III, MG 0 + 00 + 000, and MG 0 differed slightly from others with the 95th maximum percentile of 100SW being more than 28.57 g and that of SOC being more than 27.12%. Those between MG 0 + 00 + 000 and MG I+II+III, MG 0 and MG I+II+III, MG 00 and MG I+II+III, and MG 000 and MG I +II+III showed more recombination potential than the others with the 95th maximum percentile of 100SW being more than 29.99 g and that of SOC being more than 27.73%. It means that crosses from different MGs, especially those crossed with the late MGs (MG I+II+III), might have more potential in 100SW and SOC due to a broader genetic background while the crosses between the early maturity groups have less potential due to a narrowed genetic background (Table 4).

To design optimal crosses for the simultaneous improvement of 100SW and SOC, three strategies were applied. Firstly, sequential selection was applied to screen optimal crosses for simultaneous improvement in 100SW and SOC. Taking 100SW as the priority, the top 10 crosses with the highest SOC were selected from the 100 optimal crosses for 100SW, and marked as 100SW-first crosses for simultaneous improvement in 100SW and SOC (Table 5). The recombination potential of the top 10 100SW-first crosses ranged from 27.96 to 30.33 g for 100SW while that of SOC ranged from 24.79% to 25.82%. On the other hand, taking SOC as the priority, the top 10 crosses with the highest 100SW were selected from the 100 optimal crosses for SOC, and marked as the SOC-first crosses for simultaneous improvement in 100SW and SOC. The recombination potential of the top 10 SOC-first crosses ranged from 26.60% to 27.11% while that of 100SW ranged from 22.55 to 24.07 g. As for the 100SW-SOC-balance strategy, the top 10 joint crosses were selected by lining up all the predicted crosses in descending order for 100SW and SOC, respectively. The top 10 common crosses from both sides were marked as the 100SW-SOC-balance crosses for simultaneous improvement of 100SW and SOC. Its predicted recombination potential ranged from 25.48 to 26.29 g for 100SW and from 25.95% to 26.12% for SOC.

4 Discussion

4.1 Relative completeness and comparability in dissecting the QTL-allele systems of 100SW and SOC in the NECSGP through RTM-GWAS

A thorough detection of all QTLs and their multiple alleles is the key to efficient utilization of superior genes-alleles in the germplasm population in plant breeding. In this study, the QTL-allele system of 100SW and SOC in the NECSGP was detected using

the RTM-GWAS procedure. A large amount of the genetic variation of these two seed traits was explained by the detected QTLs, both large- and small-contribution QTLs, as well as their alleles. Many QTLs had a phenotype contribution of less than $R^2 < 1\%$, but altogether explained 19.51% and 16.96% phenotypic variation for 100SW and SOC in the NECSGP, respectively. Despite the high efficiency of RTM-GWAS in whole-genome QTL-allele detection, there were still unmapped genetic variations (28.27% and 16.54% for 100SW and SOC, respectively) that remained to be explored through further improvement of experiment resolution. The QTLs with a significant interaction effect with the environment were also detected in this study, but the phenotypic variation explained by the QTL $\times$ environment interaction (QEI) was relatively smaller than the main effect QTL. This may be due to the variability of the random environment factor (different years) in this study.

For comparison of QTLs detected in this study with previous studies, the worldwide information of 310 and 325 QTLs of 100SW and SOC at SoyBase (<https://www.soybase.org>) was retrieved, respectively. A total of 106 and 131 QTLs with a supporting interval of less than two centimorgans were used for comparisons with 23 and 32 QTLs overlapping to previously reported 100-SW and SOC QTLs, respectively. Previous results were mainly obtained from bi-parental populations using linkage mapping under different environments, whereas the present study utilized the NECSGP in association mapping under a uniform environment. Consequently, the identified QTL-allele system of 100SW and SOC should be relatively more comprehensive and comparable in NEC. This may be more relevant to the genetic operation of the QTL-allele systems in breeding for cultivars fitting the requirements under the environmental conditions of NEC, to design the optimal crosses in the present study.

From the above, the relative completeness and comparability in dissecting the QTL-allele systems of 100SW and SOC in the NECSGP are mainly through RTM-GWAS. This is due to the large germplasm population with broad genetic variation and the increased efficiency with two innovative procedures as indicated in the Introduction section.

4.2 Genome-wide QTL-allele dissection of a germplasm or breeding population helps the optimal utilization of complementary alleles

Recombination breeding is based on the QTL-allele or gene-allele dissection of the breeding or germplasm population, for which the establishment of a QTL-allele matrix through RTM-GWAS provided a promising approach to realize the breeders' targets. Previously, the breeders usually design their crosses based on the phenotypic values of their parental lines. Now, based on the identification of the genome-wide QTL-allele or gene-allele system, the breeders can design crosses genetically. However, sometimes the geneticists consider only individual genes with their alleles while the breeders have to consider the whole QTL-allele or gene-allele system because they do not want to leave

TABLE 5 Optimal crosses selected for simultaneous improvement of 100SW and SOC in the NECSGP.

Strategy	Order	P1				P2				Predicted progeny	
		ID	MG	100SW	SOC	ID	MG	100SW	SOC	100SW	SOC
100SW-first	1	F32	0	20.94	24.31	F35	III	20.78	20.98	28.11	25.82
	2	F406	0	24.15	22.60	P087	I	21.84	23.70	28.35	25.28
	3	F229	0	20.54	21.95	F364	0	19.89	22.98	28.33	25.00
	4	F35	III	20.78	20.98	F49	0	20.45	23.06	28.09	24.98
	5	F326	III	24.28	20.38	P087	I	21.84	23.70	28.23	24.98
	6	F58	0	23.84	22.60	P087	I	21.84	23.70	27.97	24.86
	7	F28	0	21.56	23.41	F35	III	20.78	20.98	27.96	24.85
	8	F49	0	20.45	23.06	F67	III	25.16	21.00	28.19	24.82
	9	F312	0	22.78	23.06	F35	III	20.78	20.98	28.64	24.80
	10	F326	III	24.28	20.38	F364	0	19.89	22.98	30.33	24.79
		Mean								28.42	25.02
SOC-first	1	F386	0	19.34	24.54	F49	0	20.45	23.06	24.07	26.80
	2	F386	0	19.34	24.54	F79	0	21.04	24.33	23.77	26.77
	3	F32	0	20.94	24.31	F386	0	19.34	24.54	23.48	26.91
	4	F32	0	20.94	24.31	F343	II	18.84	23.45	23.46	27.11
	5	F305	0	17.46	24.16	F32	0	20.94	24.31	23.24	26.60
	6	F306	0	17.83	24.10	F32	0	20.94	24.31	23.14	26.69
	7	F140	0	18.29	23.88	F79	0	21.04	24.33	22.91	26.65
	8	F109	II	17.29	24.00	F313	0	22.18	23.05	22.84	26.62
	9	F244	0	20.21	23.70	F343	II	18.84	23.45	22.73	26.71
	10	F343	II	18.84	23.45	F386	0	19.34	24.54	22.55	26.93
		Mean								23.22	26.78
100SW-SOC-balance	1	F309	0	22.66	23.33	F386	0	19.34	24.54	25.51	26.12
	2	F197	0	20.05	22.51	F32	0	20.94	24.31	25.48	26.09
	3	F53	0	20.58	24.19	F087	I	21.84	23.7	25.54	26.07
	4	F109	II	17.29	24.00	F35	III	20.78	20.98	25.45	26.07
	5	F140	0	18.29	23.88	F309	0	22.66	23.33	25.70	26.05
	6	F344	II	16.29	24.01	F364	0	19.89	22.98	25.63	25.99
	7	F109	II	17.29	24.00	F364	0	19.89	22.98	25.49	25.96
	8	F32	0	20.94	24.31	P087	I	21.84	23.70	26.29	25.95
	9	F306	0	17.83	24.10	P087	I	21.84	23.70	25.54	25.95
	10	F140	0	18.29	23.88	F35	III	20.78	20.98	25.96	25.95
		Mean								25.66	26.02

P1, P2, parents 1 and 2 of a cross; ID, accession name; MG, maturity group; 100SW, 100-seed weight (g); SOC, seed oil content (%); 100SW- and SOC-first, the top 10 optimal crosses selected from the 95% percentile of offspring of 100SW and SOC. 100SW-SOC-balance, the predicted respective 100SW and SOC crosses were arranged in descending order, and the top 10 joint crosses from both sides were selected as the best crosses for balanced 100SW and SOC. The parental accessions in boldface are the top 10 100SW accessions, i.e., F67, F326, F406, F147, F58, P004, P085, F36, P188, and F315 in descending order, and those in italic boldface are the top 10 SOC accessions, i.e., F82, F386, F135, F79, F32, F155, F53, F305, F306, and F351 in descending order in the NECSGP.

inferior ones in a cultivar's genetic background, especially for complex traits.

In Table 5, Supplementary Tables 5 and 6, the optimal crosses for single traits and two simultaneous traits are selected with the top 10 parents for each of the two traits marked. Not many predicted crosses with their two parents located in the top 10 accessions for each trait in the NECSGP; only a portion of predicted crosses have one parent located in the top 10 accessions for each trait in the NECSGP; even many predicted crosses do not have their one parent from the top 10 accessions for each trait in the NECSGP. This means that the complementary alleles scattered in different parental accessions. This further emphasizes the importance of whole-genome dissection for the utilization of the whole-genome complementary potential.

In addition, in the evolution from late to early maturity groups, allele emergence was quite limited whereas allele exclusion was fluctuating; this fact also supports that the evolutionary changes may be caused by the release of complementary potential due to many recombinations in history that generate phenotypic variations among maturity groups.

4.3 Genomic selection for multi-trait optimal cross design

The breeding by design concept was proposed for designing optimal crosses and selecting superior progeny, in which detecting the QTLs and their alleles was its prerequisite (Peleman and van der Voort, 2003). Breeding by design provided a basic framework for pyramiding a few major genes. Since a quantitative trait is generally controlled by numerous QTLs or genes, plant breeding breakthrough should involve in fact most QTLs and their alleles, including both large and small effect loci. Therefore, the germplasm population rather than the bi-parental population should be studied for a relatively thorough detection of the QTL-allele system. In this study, computer simulation was used to generate the progeny genotype, and the progeny genotypic value was predicted using the QTL-allele matrix. Then, optimal crosses can be selected based on progeny phenotype distribution (e.g., the 95th percentile). This *in silico* work is efficient and could not be done in the field because of the impossible large scale. The QEI QTL-allele data set can also be organized into a matrix if it is needed (Fu et al., 2020b). However, the environmental factor in the present study varied randomly, and no fixed environmental parameter was available to provide useful information in breeding for quantitative traits. Therefore, the QEI information was not used in the present analysis.

In this study, 100SW and SOC were simultaneously considered for optimal cross design. Since tight linkage between loci may exist, the two traits cannot be improved independently to achieve an ideal situation. To address this challenge, this study proposed three types of optimal crosses for the simultaneous improvement of 100SW and SOC in the NECSGP. The 100SW-first optimal crosses prioritized improving 100SW as the primary focus, with SOC being considered as the secondary objective. In contrast, the SOC-first optimal crosses focused on SOC at first and then 100SW. In addition, 100SW-SOC-balance optimal crosses focused not on the extreme

value of one trait, but the best overall. Instead of separately focusing on the extreme values of one trait, this study aimed to also identify optimal crosses that would result in the best overall improvement of both 100SW and SOC. Here, the top 2,000 crosses were used for ordering in each trait. It depends on the number cases, and 2,000 crosses for each trait are enough in this study. Based on these three types of optimal crosses, this study presented a preliminary attempt at multi-trait cross design in the NECSGP. The improvement of 100SW and SOC responding to the three simultaneous selection strategies may raise both traits' levels, but different from each other. The researchers can choose their preferred one based on their requirements. For example, the SOC-first strategy may be considered because SOC is economically more important than 100SW. In addition, the method can be extended to simultaneous selection for more traits. However, it is important to note that the methods based on a composite selection index may offer more sophisticated approaches for multi-trait selection in breeding.

The concept of GS using genome-wide markers was introduced as a method of predicting breeding values of a group of traits. This approach, proposed by Meuwissen et al. (2001), aimed to leverage the information provided by the entire genome rather than focusing on specific loci of some specific traits, especially in animal breeding. In classical GS, it is assumed that all markers across the genome have effects on the trait of interest. These effects are then estimated using a reference or training population, which serves as the basis for establishing a prediction model. By considering the collective influence of multiple markers, GS offers the potential to capture the polygenic nature of many complex traits. However, it is important to note that genetic mechanisms governing different quantitative traits can vary significantly. The complexity of these mechanisms can also differ, making it challenging to apply a uniform approach to all traits. Therefore, it becomes crucial to consider the specific characteristics of the trait under consideration when implementing GS.

The use of QTL-allele matrices allows for a more targeted and tailored approach to GS. By focusing on the specific alleles associated with QTLs, breeders can gain insights into the genetic basis of the trait and make informed decisions regarding optimal crosses. This approach acknowledges the unique characteristics and complexities of different quantitative traits, offering a more suitable framework for designing breeding strategies in plants. Therefore, genome-wide breeding by design in fact is the GS based on the genome-wide QTL-allele (or gene-allele) matrix, which should be more appropriate in designing optimal crosses in plants.

Data availability statement

The data presented in the study are deposited in the GitHub repository, accession number: <https://github.com/njau-sri/NECSGP-100SW-SOC>.

Author contributions

JH: Methodology, Investigation, Formal analysis, Writing – review & editing, Writing – original draft. LF: Investigation,

Writing – review & editing, Writing – original draft. XH: Formal analysis, Writing – review & editing, Writing – original draft. YCW: Writing – review & editing, Writing – original draft, Formal analysis. MW: Formal analysis, Writing – review & editing, Writing – original draft. QZ: Formal analysis, Writing – review & editing, Writing – original draft. WF: Formal analysis, Writing – review & editing, Writing – original draft. MF: Investigation, Writing – review & editing, Writing – original draft. YPW: Investigation, Writing – review & editing, Writing – original draft. HR: Investigation, Writing – review & editing, Writing – original draft. WD: Investigation, Writing – review & editing, Writing – original draft. WW: Investigation, Writing – review & editing, Writing – original draft. JG: Supervision, Methodology, Writing – review & editing, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Science and Technology Innovation 2030-Major Project (2023ZD0403401), the Zhongshan Biological Breeding Laboratory (ZSBBL-KY2023-03), the National Key Research and Development Program of China (2021YFF1001204 and 2021YFD1201602), the MOE 111 Project (B08025), the Fundamental Research Funds for the Central Universities (KYZZ201901), the MARA CARS-04 program, the Primary

Research & Development Plan of Jiangsu Province (BE2021358), the Guidance Foundation of Sanya Institute of Nanjing Agricultural University (NAUSY-ZZ02, NAUSY-MS05), and the Bioinformatics Center of Nanjing Agricultural University.

Conflict of interest

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

Publisher's note

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.2024.1483995/full#supplementary-material>

References

- Andolfatto, P., Davison, D., Erezylmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., et al. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21, 610–617. doi: 10.1101/gr.115402.110
- Bandillo, N., Jarquin, D., Song, Q. J., Nelson, R., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome* 8, plantgenome2015.04.0024. doi: 10.3835/plantgenome2015.04.0024
- Chen, X., Liu, C., Guo, P., Hao, X., Pan, Y., Zhang, K., et al. (2023). Differential SW16.1 allelic effects and genetic backgrounds contributed to increased seed weight after soybean domestication. *J. Integr. Plant Biol.* 65, 1734–1752. doi: 10.1111/jipb.13480
- Collard, B. C. Y., and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Phil. Trans. R. Soc. B* 363, 557–572. doi: 10.1098/rstb.2007.2170
- Fahim, A. M., Liu, F., He, J., Wang, W., Xing, G., and Gai, J. (2021). Evolutionary QTL-allele changes in main stem node number among geographic and seasonal subpopulations of Chinese cultivated soybeans. *Mol. Genet. Genomics* 296, 313–330. doi: 10.1007/s00438-020-01748-9
- Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. *Genome Biol.* 18, 161. doi: 10.1186/s13059-017-1289-9
- Fasoula, V. A., Harris, D. K., and Boerma, H. R. (2004). Validation and designation of quantitative trait loci for seed protein, seed oil, and seed weight from two soybean populations. *Crop Sci.* 44, 1218–1225. doi: 10.2135/cropsci2004.1218
- Feng, W., Fu, L., Fu, M., Sang, Z., Wang, Y., Wang, L., et al. (2022). Transgressive potential prediction and optimal cross design of seed protein content in the northeast China soybean population based on full exploration of the QTL-allele system. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.896549
- Fu, M., Wang, Y., Ren, H., Du, W., Wang, D., Bao, R., et al. (2020a). Genetic dynamics of earlier maturity group emergence in south-to-north extension of Northeast China soybeans. *Theor. Appl. Genet.* 133, 1839–1857. doi: 10.1007/s00122-020-03558-4
- Fu, M., Wang, Y., Ren, H., Du, W., Yang, X., Wang, D., et al. (2020b). Exploring the QTL-allele constitution of main stem node number and its differentiation among maturity groups in a Northeast China soybean population. *Crop Sci.* 60, 1223–1238. doi: 10.1002/csc2.20024
- Guo, B., Sun, L., Ren, H., Sun, R., Wei, Z., Hong, H., et al. (2022). Soybean genetic resources contributing to sustainable protein production. *Theor. Appl. Genet.* 135, 4095–4121. doi: 10.1007/s00122-022-04222-9
- Hanson, C. H., Robinson, H. F., and Comstock, R. E. (1956). Biometrical studies of yield in segregating populations of Korean lespedeza. *Agron. J.* 48, 268–272. doi: 10.2134/agronj1956.00021962004800060008x
- He, J., and Gai, J. (2020). QTL-allele matrix detected from RTM-GWAS is a powerful tool for studies in genetics, evolution, and breeding by design of crops. *J. Integr. Agric.* 19, 1407–1410. doi: 10.1016/S2095-3119(20)63199
- He, J., Meng, S., Zhao, T., Xing, G., Yang, S., Li, Y., et al. (2017). An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding. *Theor. Appl. Genet.* 130, 2327–2343. doi: 10.1007/s00122-017-2962-9
- Hong, J., Su, S., Wang, L., Bai, S., Xu, J., Li, Z., et al. (2023). Combined genome-wide association study and epistasis analysis reveal multifaceted genetic architectures of plant height in Asian cultivated rice. *Plant Cell Environ.* 46, 1295–1311. doi: 10.1111/pce.14557
- Hwang, E. Y., Song, Q., Jia, G., Specht, J. E., Hyten, D. L., Costa, J., et al. (2014). A genome-wide association study of seed protein and oil content in soybean. *BMC Genomics* 15, 1. doi: 10.1186/1471-2164-15-1
- Khan, M. A., Tong, F., Wang, W., He, J., Zhao, T., and Gai, J. (2018). Analysis of QTL-allele system conferring drought tolerance at seedling stage in a nested association mapping population of soybean [Glycine max (L.) Merr.] using a novel GWAS procedure. *Planta* 248, 947–962. doi: 10.1007/s00425-018-2952-4
- Khatun, M., Monir, M. M., Lou, X., Zhu, J., and Xu, H. (2022). Genome-wide association studies revealed complex genetic architecture and breeding perspective of maize ear traits. *BMC Plant Biol.* 22, 537. doi: 10.1186/s12870-022-03913-1
- Li, S., Cao, Y., He, J., Zhao, T., and Gai, J. (2017). Detecting the QTL-allele system conferring flowering date in a nested association mapping population of soybean using a novel procedure. *Theor. Appl. Genet.* 130, 2297–2314. doi: 10.1007/s00122-017-2960-y
- Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* 10, 387–406. doi: 10.1146/annurev.genom.9.081307.164242

- Li, Y.-H., Qin, C., Wang, L., Jiao, C., Hong, H., Tian, Y., et al. (2022). Genome-wide signatures of the geographic expansion and breeding of soybean. *Sci. China Life Sci.* 66, 350–365. doi: 10.1007/s11427-022-2158-7
- Liu, X., He, J., Wang, Y., Xing, G., Li, Y., Yang, S., et al. (2020). Geographic differentiation and phylogeographic relationships among world soybean populations. *Crop J.* 8, 260–272. doi: 10.1016/j.cj.2019.09.010
- Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819
- Miao, L., Yang, S., Zhang, K., He, J., Wu, C., Ren, Y., et al. (2020). Natural variation and selection in GmSWEET39 affect soybean seed oil content. *New Phytol.* 225, 1651–1666. doi: 10.1111/nph.16250
- Murray, M. G., and Thompson, W. F. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* 8, 4321–4326. doi: 10.1093/nar/8.19.4321
- Nichols, D. M., Glover, K. D., Carlson, S. R., Specht, J. E., and Diers, B. W. (2006). Fine mapping of a seed protein qtl on soybean linkage group i and its correlated effects on agronomic traits. *Crop Sci.* 46, 834–839. doi: 10.2135/cropsci2005.05-0168
- Pathan, S. M., Vuong, T., Clark, K., Lee, J., Shannon, J. G., Roberts, C. A., et al. (2013). Genetic mapping and confirmation of quantitative trait loci for seed protein and oil contents and seed weight in soybean. *Crop Sci.* 53, 765–774. doi: 10.2135/cropsci2012.03.0153
- Peleman, J. D., and van der Voort, J. R. (2003). Breeding by design. *Trends Plant Sci.* 8, 330–334. doi: 10.1016/S1360-1385(03)00134-1
- Ray, S., Jarquin, D., and Howard, R. (2022). Comparing artificial-intelligence techniques with state-of-the-art parametric prediction models for predicting soybean traits. *Plant Genome* n/a, e20263. doi: 10.1002/tpg2.20263
- Ren, W., Zhao, L., Liang, J., Wang, L., Chen, L., Li, P., et al. (2022). Genome-wide dissection of changes in maize root system architecture during modern breeding. *Nat. Plants* 8, 1408–1422. doi: 10.1038/s41477-022-01274-z
- Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802
- 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
- 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
- Su, Y., Zhang, Z., He, J., Zeng, W., Cai, Z., Lai, Z., et al. (2023). Gene-allele system of shade tolerance in southern China soybean germplasm revealed by genome-wide association study using gene-allele sequence as markers. *Theor. Appl. Genet.* 136, 152. doi: 10.1007/s00122-023-04390-2
- Ullah Zaid, I., Tang, W., He, J., Ullah Khan, S., and Hong, D. (2018). Association analysis uncovers the genetic basis of general combining ability of 11 yield-related traits in parents of hybrid rice. *AoB Plants* 11, ply077. doi: 10.1093/aobpla/ply077
- Wang, L., Liu, F., Hao, X., Wang, W., Xing, G., Luo, J., et al. (2021). Identification of the QTL-allele system underlying two high-throughput physiological traits in the Chinese soybean germplasm population. *Front. Genet.* 12. doi: 10.3389/fgene.2021.600444
- Wang, S., Liu, S., Wang, J., Yokosho, K., Zhou, B., Yu, Y. C., et al. (2020). Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. *Natl. Sci. Rev.* 7, 1776–1786. doi: 10.1093/nsr/nwaa110
- Wang, W., Zhou, B., He, J., Zhao, J., Liu, C., Chen, X., et al. (2020). Comprehensive identification of drought tolerance QTL-allele and candidate gene systems in Chinese cultivated soybean population. *Int. J. Mol. Sci.* 21, 4830. doi: 10.3390/ijms21144830
- Yang, H., Wang, W., He, Q., Xiang, S., Tian, D., Zhao, T., et al. (2019). Identifying a wild allele conferring small seed size, high protein content and low oil content using chromosome segment substitution lines in soybean. *Theor. Appl. Genet.* 132, 2793–2807. doi: 10.1007/s00122-019-03388-z
- Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X., Pool, J. E., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329, 75–78. doi: 10.1126/science.1190371
- Zhang, Y., He, J., Wang, H., Meng, S., Xing, G., Li, Y., et al. (2018). Detecting the QTL-allele system of seed oil traits using multi-locus genome-wide association analysis for population characterization and optimal cross prediction in soybean. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.01793
- Zhang, Y., He, J., Wang, Y., Xing, G., Zhao, J., Li, Y., et al. (2015). Establishment of a 100-seed weight quantitative trait locus-allele matrix of the germplasm population for optimal recombination design in soybean breeding programmes. *J. Exp. Bot.* 66, 6311–6325. doi: 10.1093/jxb/erv342



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Kaushal Pratap Singh,
Directorate of Rapeseed Mustard Research
(DRMR), India
Senguttuvel P.,
Indian Institute of Rice Research (ICAR), India

*CORRESPONDENCE

Parveen Chhuneja
✉ pchhuneja@pau.edu

†PRESENT ADDRESS

Navaneetha Krishnan J.,
Department of Biotechnology, Pachhunga
University College Campus, Mizoram
University, Aizawl, India
Guriqbal Singh Dhillon,
College of Agriculture and Life Sciences,
University of Idaho, Aberdeen, Idaho,
United States

RECEIVED 25 August 2024

ACCEPTED 27 November 2024

PUBLISHED 20 December 2024

CITATION

Krishnan J. N., Kaur S., Kumar U., Singh R.,
Dhillon GS, Bhati PK and Chhuneja P (2024)
Mapping heat tolerance QTLs in *Triticum*
durum-*Aegilops speltoides* backcross
introgression lines to enhance
thermotolerance in wheat.
Front. Plant Sci. 15:1485914.
doi: 10.3389/fpls.2024.1485914

COPYRIGHT

© 2024 Krishnan J., Kaur, Kumar, Singh,
Dhillon, Bhati and Chhuneja. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Mapping heat tolerance QTLs in *Triticum durum*-*Aegilops speltoides* backcross introgression lines to enhance thermotolerance in wheat

Navaneetha Krishnan J.^{1†}, Satinder Kaur¹, Uttam Kumar^{2,3},
Rohtas Singh¹, Guriqbal Singh Dhillon^{1†}, Pradeep Kumar Bhati²
and Parveen Chhuneja^{1*}

¹School of Agricultural Biotechnology, Punjab Agricultural University, Ludhiana, India, ²Borlaug
Institute for South Asia, Ludhiana, India, ³Astralyn Agro One Person Company (OPC) Pvt. Ltd,
Shamli, India

Wheat, a major cereal crop, is the most consumed staple food after rice in India. Frequent episodes of heat waves during the past decade have raised concerns about food security under impending global warming and necessitate the development of heat-tolerant wheat cultivars. Wild relatives of crop plants serve as untapped reservoirs of novel genetic variations. In the present study a mapping population comprising 311 BC₂F₁₀ backcross introgression lines (BILs) developed by crossing *Triticum durum* and heat-tolerant diploid wild wheat relative *Aegilops speltoides* accession pau3809 was used to map QTLs for terminal heat tolerance. The homozygous BILs were evaluated for heat stress tolerance component traits under an optimum environment (OE) and a heat-stressed environment (HE) for the two cropping seasons. Data on spike length, spikelet number per spike, peduncle length, thousand-grain weight, grains per spike, days to heading, days to maturity, grain filling duration, NDVI at heading, plant height and plot yield were recorded. Genotyping-by-sequencing (GBS) of the BILs was carried out, and 2945 high-quality, polymorphic SNPs were obtained. Thirty QTLs were detected for various heat tolerance component traits on chromosomes 1A, 1B, 2A, 2B, 3B, 4B, 5A, 5B, 6A and 6B with phenotypic variance ranging from 5 to 11.5%. Several candidate genes reported to play a role in heat stress responses were identified by browsing the 1.85 Mb physical region flanking the stable QTLs detected under the HE. Identified QTL and linked markers can be employed for genomics-assisted breeding for heat tolerance in wheat.

KEYWORDS

heat tolerance, wheat, *Aegilops speltoides*, QTL mapping, *Triticum durum*, backcross introgression lines

1 Introduction

Wheat is a major staple food cultivated across the globe and provides 20% of the dietary energy and protein for global population (Curtis et al., 2002; Shiferaw et al., 2013). The developing and developed countries contribute equally to global wheat production (Shiferaw et al., 2013). After rice, wheat is the predominant staple diet of the Indian population and is the chief ingredient of a variety of processed foods (Khatkar et al., 2016). India occupies second rank among the wheat-producing countries with a production of 110.5 million tonnes and a productivity of 3.47 tonnes/ha (Government of India (GOI), 2023). Wheat production is under constant threat due to several factors including decreasing water resources, soil fertility loss, emerging pests and diseases and global warming (Ramadas et al., 2020; Asseng et al., 2015).

Climate change poses significant risks to wheat production in the form of frequent episodes of heat waves, drought and expansion of soil salinity (Joshi and Kar, 2009; Tripathi et al., 2016; Mondal et al., 2021). High-temperature stress is emerging as a serious threat to wheat productivity and affects nearly 40% of the irrigated wheat crop globally (Iqbal et al., 2017; Farhad et al., 2023). Heat stress at the reproductive stage impairs grain-filling and seed set, resulting in reduced grain yield (Hays et al., 2007). The Indo-Gangetic plains zone, which contributes 15% to global wheat production, is more prone to climatic shifts, and it is predicted that by the end of 2050, 51% of its total area will be reclassified as a heat-stressed environment with a shortened growth season (Ortiz et al., 2008). Lobell et al. (2011) reported that during the 1980-2006 period, wheat yields were reduced up to 5.5% as a result of rising temperatures. Various metabolic processes are implicated during heat stress resulting in decreased grain weight, shriveled grains, early senescence, altered starch-lipid composition in grains, reduced starch accumulation, reduced seed germination and loss of vigour (Balla et al., 2012). Therefore, breeding for heat tolerance has gained priority in the national and international wheat improvement programmes. Heat tolerance *per se* is not a Mendelian trait, as it comprises different individual components that interact with each other to manifest heat tolerance. Several morphological, physiological, biochemical and yield-related traits are involved in conferring or expressing heat tolerance in a plant. In the case of wheat, these include and are not limited to stay green, biomass, canopy temperature, membrane thermostability, early ground cover, grain weight, grain number, tiller number, photosynthetic efficiency, production of reactive oxygen species (ROS), antioxidant enzymes and chlorophyll content (Sundeep et al., 2013; Asthir, 2015; Narayanan, 2018).

Developing breeding populations utilizing diverse genetic resources and characterizing them both phenotypically and genotypically can lead to the delineation of the genomic regions (QTL) associated with heat tolerance (Kim et al., 2017; Peverelli and Rogers, 2013). With the availability of the whole genome sequence of wheat and the recent expansion in genomics and bioinformatics tools, wheat breeding has increased its pace (Appels et al., 2018; Rasheed and Xia, 2019; Adamski et al., 2020). QTLs for heat tolerance-related traits have been reported in all 21 wheat

chromosomes (Bhusal et al., 2017). There are limited reports of mapping QTLs introgressed from wild species for heat stress tolerance in wheat (Awlachev et al., 2016).

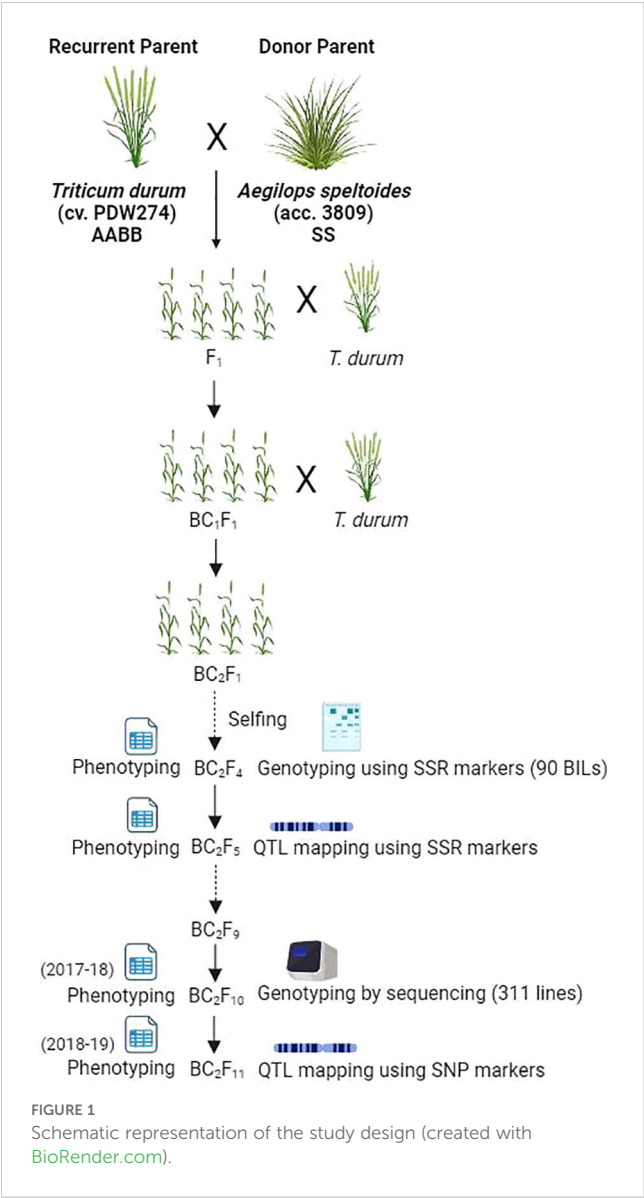
The genetic diversity for heat stress tolerance in the cultivated wheat germplasm is limited. Wild species of wheat have immensely contributed to wheat improvement in terms of being reservoirs of genes/QTLs for various biotic and abiotic stresses (Gill et al., 2006; Schneider et al., 2008; Pradhan et al., 2012; Kishii, 2019; Crespo-Herrera et al., 2019). *Aegilops speltoides* has been identified as a promising germplasm source for heat stress tolerance traits (Ehdaie and Waines, 1992; Waines, 1994; Khanna-Chopra and Viswanathan, 1999; Pradhan et al., 2012; Awlachev et al., 2016). Homeologous pairing in wheat is inhibited by the *Ph1* locus present in the long arm of the 5B chromosome (Okamoto, 1957; Sears and Okamoto, 1958; Riley and Chapman, 1958; Bhullar et al., 2014). *Ae. speltoides*, the probable donor of the wheat B genome, has genes that are epistatic to the *Ph1* locus, and these genes, called *Ph1* suppressors, were mapped to the 3S and 7S chromosomes by Dvorak et al. (2006). The degree of suppression varies among the different accessions of *Ae. speltoides*, and accordingly, they are categorized as strong, moderate or weak suppressors of the *Ph1* locus (Millet, 2008). Accessions of *Ae. speltoides* having moderate suppression activity of the *Ph1* locus can aid in the transfer of useful variability to cultivated wheat with minimal meiotic anomalies (Millet, 2008).

We had earlier reported the development of a set of 90 tetraploid genome-wide homozygous backcross introgression lines involving *T. durum* cv. PDW274 and a heat-tolerant accession of *Ae. speltoides* (Awlachev et al., 2016). We carried out introgression profiling of the BILs with 152 polymorphic SSR markers and mapped QTLs for heat tolerance component traits (Awlachev et al., 2016). For deep diving into the genome of *T. durum*-*Ae. speltoides*, backcross introgression lines for heat tolerance genes/QTL, we expanded the introgression panel and performed genotyping-by-sequencing to increase the resolution of QTL mapping. Identification and mapping of heat tolerance QTL using genome-wide SNP markers is being discussed in the present manuscript.

2 Materials and methods

2.1 Development of the plant material

A heat-tolerant accession of *Ae. speltoides* (2n=2x=14, SS) designated as accession pau3809 was crossed with *T. durum* cv. PDW274 (2n=4x=28, AABB) as the female parent. The F₁ plants displayed profuse tillering and were extensively backcrossed with PDW274. Five BC₁F₁ plants were obtained after pollinating more than 1500 florets. The five BC₁F₁ plants were again backcrossed with the female parent to generate 128 BC₂F₁s (Awlachev et al., 2016). All the BC₂F₁s were selfed, and a single seed descent strategy was employed to develop 311 BC₂F₁₀ homozygous BILs. The crossing strategy adopted for generating the plant material is schematically depicted in Figure 1.



2.2 Field evaluation for heat stress tolerance

The BILs and the recurrent parent PDW274 were evaluated at the experimental farm, Borlaug Institute for South Asia (BISA),

Ladhowal, Punjab, for two consecutive crop seasons(2017-18 and 2018-19). The BILs were sown on two different dates: first during the third week of November, referred to as the optimum environment (OE) and second, during the third week of December, referred to as the heat-stressed environment (HE) in a randomized block design with two replications. The OE and HE of the 2017-18 cropping season will be referred to as OE1 and HE1, and the OE and HE of the 2018-19 cropping season will be referred to as OE2 and HE2 in the subsequent sections. The details of the sowing dates are given in Table 1. Sowing was carried out using a limit plot planter in four-row plots of size 2.6 x 0.88m. Standard crop husbandry practices were followed to ensure a healthy crop. An aerial view of the experimental field trials is shown in Supplementary Figure S1. The Field Book app, an Android-based open-source application developed by Rife and Poland (2014), was used for recording the phenotypic data of all the traits except plot yield, thousand-grain weight and grains per spike. Since we delayed the sowing by a month in the case of the HE trials, we believe that the BILs experienced significant exposure to heat stress at the reproductive stage, except for very few lines that flowered too early.

Data on the following phenotypic traits were recorded during the two cropping seasons: Days to heading (DTH) was calculated on a plot basis as the number of days required for the completion of spike exertion [Zadoks' growth stage (GS) 58] of at least 50% of the plot's vegetation cover. Days to maturity (DTM) was calculated on a plot basis as the number of days required for attaining physiological maturity (GS91) of at least 75% of the plot's vegetation cover. For grain-filling duration (GFD), the period between the days to heading (GS58) and the completion of physiological maturity (GS91) was noted and expressed in days. Five mature primary spikes were selected at random from each plot, and the mean spikelet number per spike was calculated and recorded as spikelet number per spike (SN). The selected spikes were threshed, and the average count of the grains was recorded as grain number per spike (GPS). Spike length (SL) was calculated as the average length of five randomly selected mature primary spikes from each plot. For peduncle length (PL), the average length of peduncles corresponding to five randomly selected primary spikes was measured at physiological maturity. Plot yield (PY) was recorded as the grain yield of individual plots. Normalized difference vegetation index at heading (NDVI_H) was measured using a handheld 'Greenseeker' device (Trimble, USA) at the heading stage (GS 58). One reading per plot was recorded by passing the

TABLE 1 Sowing dates, harvesting dates, temperature and rainfall data for the two-year trials of the *T. durum*-*Ae. speltooides* backcross introgression lines.

Year	Sowing date		Harvesting date		Mean maximum temperature at grain filling stage		Mean minimum temperature at grain filling stage		Total rainfall received (mm)
Environment	Optimum	Heat-stressed	Optimum	Heat-stressed	Optimum	Heat-stressed	Optimum	Heat-stressed	Across the crop season
2017-18	23-11-2017	20-12-2017	29-04-2018	09-05-2018	30.5	33.7	15.6	17.7	125.4
2018-19	16-11-2018	20-12-2018	06-05-2019	15-05-2019	28.0	32.7	13.7	17.2	253.9

Greenseeker at 50–60 cm above the plot surface for 4–5 seconds. The device was passed over the middle of the plot canopy, excluding the exterior rows. The thousand-grain weight (TGW) was estimated using a novel image analysis-based method reported in our earlier publication (Krishnan et al., 2023).

2.3 Weather data

Weather data was recorded for the entire wheat growing season of 2017–18 and 2018–19 using the automatic weather data recorder at BISA, Ladhawal. Data on daily minimum and maximum temperature and rainfall were collected.

2.4 Statistical analysis of the phenotypic data

The best linear unbiased predictors (BLUPs) of the phenotypic values were calculated for the optimum and heat-stressed environments of both cropping seasons using the META-R software (Alvarado et al., 2020). The BLUPs were calculated for individual years and across the years for both environments, respectively. Summary statistics were computed from the BLUPs using the PAST 4.03 software (Hammer et al., 2001). The trait distribution, plotting of weather data, and correlation analysis were carried out using the 'ggplot2' and 'corrplot' packages in the R software (R Core Team, 2013; Wickham, 2016; Wei and Simko, 2021). For correlation analysis, BLUPs of the phenotypic traits over the years (pooled data) were used. Levene's test was used to test for the equality of variances for the year-wise phenotypic data using the 'Proc GLM' procedure in SAS 9.4 (SAS Institute Inc, 2013). The analysis of variance for the two-year phenotypic data was carried out using SAS 9.4 (SAS Institute Inc, 2013).

2.5 Genotyping and introgression profiling

DNA was extracted from the BILs, the recurrent parent PDW274 and the donor parent *Ae. speltoides* acc. pau3809 using the protocol developed by Allen et al. (2006). The sequencing of the BILs was carried out using the two-enzyme genotyping-by-sequencing (GBS) approach described by Poland et al. (2012). The raw sequence files were processed using the GBS pipeline version 5.2.31 in the TASSEL software (Glaubitz et al., 2014). The SNP calling was done against the A and B genomes of the wheat genome assembly Refseq v1.0. The generated vcf file was filtered for a minimum read depth of 3 (DP3) and converted into hapmap format, resulting in 1,15,709 SNPs. The hapmap file was further filtered to identify homozygous SNPs for each parental line, and SNPs that were polymorphic between the two parents were retained that reduced the number to 10,092 SNPs. Loci with very low coverage (<50%) and high heterozygosity (>20%) were filtered out, and BILs containing more than 10% missing data were excluded. Finally, we ended up with 227 BILs with high-quality SNP data, comprising 2945 polymorphic SNPs that was used for QTL mapping. Introgression profiling was carried out

using the GGT2 software (Van Berloo, 2008) to visualize the donor introgressions.

2.6 QTL mapping

The BLUPs of the two-year phenotypic data, pooled phenotypic data and genotypic data of the 227 BILs and the recurrent parent were used for QTL mapping in QTL IciMapping version 4.2 (Meng et al., 2015). The mapping was carried out using the likelihood ratio test based on stepwise regression for the additive QTL (RSTEP-LRT-ADD) method. The logarithm of odds (LOD) significance was fixed at 2.5. QTLs with a LOD greater than 2.5 and phenotypic variance explained (PVE) above 5% were considered significant.

2.7 QTL nomenclature

QTL names were assigned based on the guidelines of the International Rules of Genetic Nomenclature (Boden et al., 2023). QTLs were named in the following manner: "Q" denotes "QTL", followed by abbreviations of the phenotypic traits (PY, TGW, GPS, SN, SL, PL, DTH, DTM, GFD and NDVI_H); "pau" stands for Punjab Agricultural University; "Td" denotes that the QTL donor parent is *T. durum*; and "As" denotes that the QTL is donated by *Ae. speltoides*; "OE" and "HE" are the abbreviations for optimum and heat-stressed environments; the last is the wheat chromosome number; if there are more than one QTL detected on a chromosome, it is numbered as 1,2,3, etc., based on their physical position.

2.8 Testing the allelic effect of the SNPs associated with the stable QTLs

The effect of the allelic substitution at the SNP locus associated with the stable QTLs detected in the heat-stressed environment was evaluated using the Mann-Whitney U test. The genotypes were divided into two groups based on their SNP allelic pattern. The pooled BLUPs of the two-year phenotypic data were used to generate the box plots, and the two groups were compared to test for significant median differences following the Mann-Whitney U test. Since the phenotypic data of the two allelic groups were not normally distributed, we could not proceed with the parametric tests for mean comparison. The box plots and statistics were generated using the 'ggbetweenstats' function of the 'ggstatsplot' R-package (Patil, 2021).

2.9 Postulation of candidate genes

The pairwise linkage disequilibrium between the SNP loci at the genome level with no missing data was calculated in Tassel v 5.0 and plotted by computing the r^2 estimators between all pairs of SNP markers using the R software. LD decay was estimated using a spline that was fitted on a $LD_{\frac{1}{2},95}$, the distance at which the short-range LD

is halved when using the 95% percentile of r^2 at short range, as an estimator for LD. The chromosome-wise linkage disequilibrium (LD) decay value (in Mb) was used to define the confidence intervals to locate candidate genes. The physical interval flanking the consistent QTLs was mined for putative candidate genes based on the list of high-confidence genes in the IWGSC RefSeq v1.0 genome assembly. The web-based genome browser ‘Persephonesoft’ (<https://web.persephonesoft.com/>) was used for fetching the gene IDs of high-confidence genes. The functional annotations of the high-confidence genes were retrieved from https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/.

3 Results

3.1 Phenotyping for heat stress tolerance component traits

The summary statistics of the heat tolerance component traits are enlisted in Table 2, and it is evident from the mean statistics that heat stress had a negative effect on the agronomic performance of the BILs in both cropping seasons. The BILs displayed considerable variation

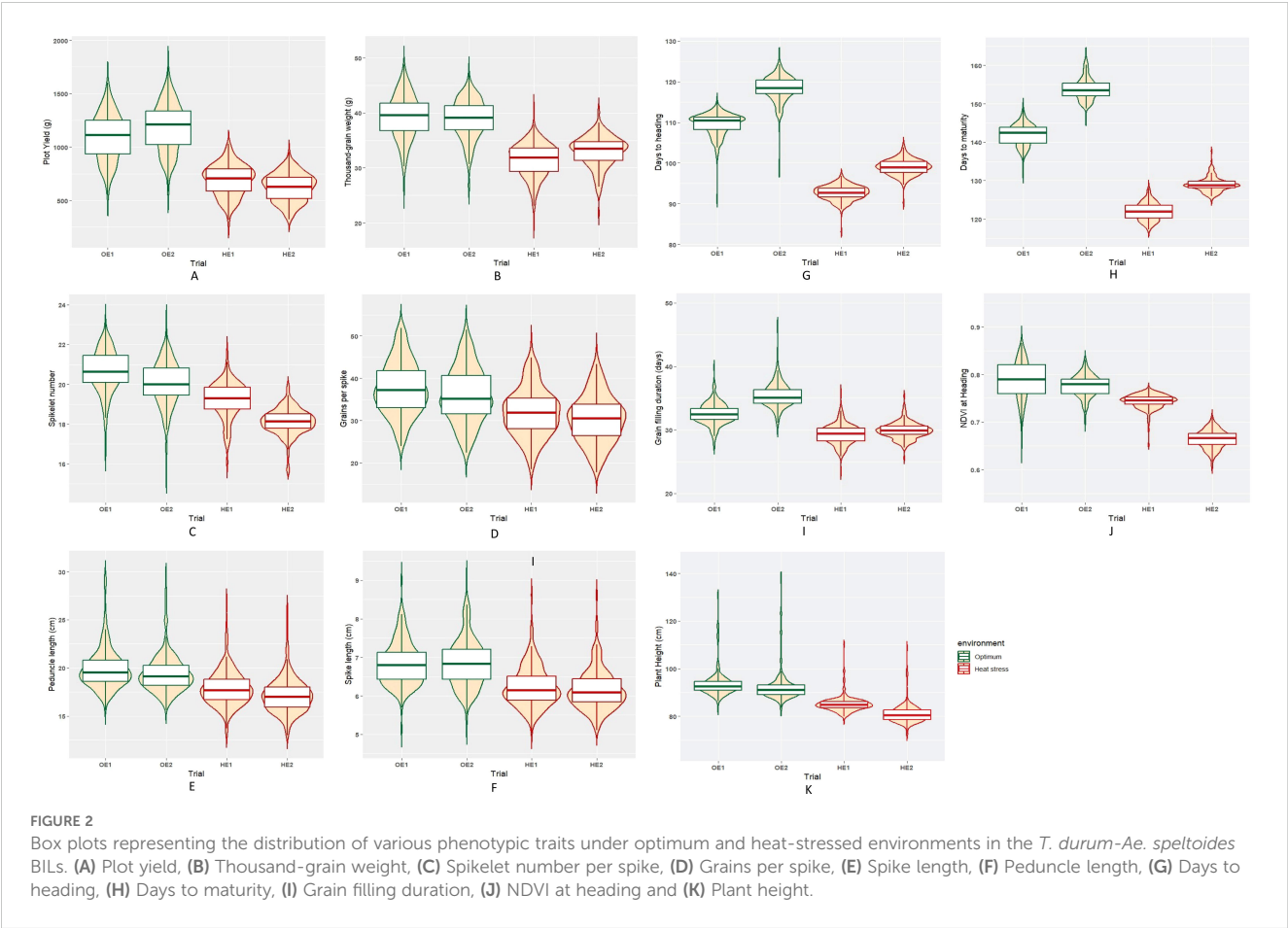
for all the traits studied under both OE and HE and the relative distribution of various phenotypic traits under both environments is explained graphically in the form of boxplots overlaid with violin plots (Figure 2). The variations observed for spike and peduncle traits among the BILs and the parents are presented in Supplementary Figures S2, S3, respectively. Regarding the temperature data during the grain filling period, there was a difference of 3° to 4°C in the mean maximum and minimum temperatures between the optimum and heat-stressed environment trials (Table 1; Supplementary Figure S4). This validates the incidence of heat stress during the reproductive phase of the HE trial, and the delay of the sowing date was effective in imposing terminal heat stress. In the HE2 trial, there was a greater decline in the plot yield and NDVI of the BILs in comparison to the HE1 trial due to the high rainfall received during the January and February months (Supplementary Figure S5). As the late-sown BILs were in the early vegetative stage, there was a reduction in their growth and development due to waterlogging stress. The plant density of many plots declined as some proportion of the plants succumbed to the stress, leading to a decrease in the plot yield and NDVI.

Levene’s test for the homogeneity of variances was carried out using the mean values of the phenotypic traits recorded during the

TABLE 2 Summary statistics for the optimum and heat-stressed environment trials of *T. durum*-*Ae. speltoides* backcross introgression lines.

Statistics	Environment	PY (g)	TGW (g)	GPS	SN	SL (cm)	PL (cm)	DTH (days)	DTM (days)	GFD (days)	NDVI_H	PH (cm)
Mean	OE1	1098.01	39.33	37.58	20.60	6.83	19.93	109.55	142.04	32.49	0.79	93.54
	HE1	691.91	31.48	31.87	19.24	6.23	17.87	92.65	122.02	29.37	0.74	85.05
	OE2	1186.53	38.88	36.01	20.03	6.85	19.43	118.38	153.86	35.47	0.78	92.39
	HE2	621.93	33.01	30.30	18.13	6.15	17.16	99.02	129.04	30.02	0.66	81.00
Minimum	OE1	522.00	25.26	23.45	15.67	4.95	15.35	90.00	131.00	26.50	0.65	82.97
	HE1	246.50	19.29	17.90	15.35	4.90	12.85	81.50	116.50	22.00	0.48	78.13
	OE2	552.00	25.45	22.00	14.33	5.07	15.35	98.00	146.00	29.50	0.70	82.68
	HE2	287.00	21.12	17.30	14.67	4.98	12.75	89.00	124.00	21.50	0.52	72.20
Maximum	OE1	1629.00	49.26	52.40	23.67	9.25	30.25	115.50	149.50	41.00	0.87	130.76
	HE1	1053.00	41.14	48.35	22.18	8.90	27.32	98.00	129.00	37.50	0.83	110.15
	OE2	1773.50	47.89	51.90	23.83	9.25	30.15	126.50	163.00	48.00	0.83	138.10
	HE2	986.50	40.90	46.25	20.67	8.92	26.67	105.50	137.00	38.00	0.78	108.71
Standard deviation	OE1	225.65	4.16	6.44	1.33	0.61	2.27	2.96	3.05	2.00	0.04	5.68
	HE1	155.09	3.21	5.91	1.09	0.58	1.93	2.15	2.66	1.98	0.04	3.21
	OE2	235.20	3.73	6.64	1.41	0.69	2.15	3.20	2.99	2.21	0.02	6.72
	HE2	143.47	2.98	5.86	0.93	0.60	1.99	2.19	2.42	2.18	0.05	4.09
Coefficient of Variation	OE1	20.55	10.57	17.13	6.45	8.95	11.38	2.70	2.14	6.15	5.12	6.07
	HE1	22.41	10.19	18.53	5.65	9.26	10.80	2.32	2.18	6.73	5.73	3.77
	OE2	19.82	9.60	18.42	7.03	10.09	11.08	2.71	1.95	6.24	2.95	7.27
	HE2	23.06	9.03	19.35	5.13	9.50	11.60	2.18	1.87	7.25	7.32	5.05

PY, Plot yield; TGW, Thousand-grain weight; GPS, Grains per spike; SN, Spikelet number per spike; SL, Spike length; PL, Peduncle length; DTH, Days to heading; DTM, Days to maturity; GFD, Grain filling duration; NDVI_H, NDVI at heading; PH, Plant height.



two seasons (years) for OE and HE trials. The results of Levene’s test revealed that the variances were equal for all the traits of both OE and HE trials, except NDVI_H in the case of the OE trial (Supplementary Table S1). Since the variances were homogeneous across the years, pooled analysis of variance was carried out for the optimum and heat-stressed trials (Tables 3, 4). Under both environments, the genotype effects were significant for all the traits studied, whereas the effects of years and genotype by year were significant for most of the traits. There was notable variation in the temperature and rainfall data among the two cropping seasons (Table 1). Due to the extended winter during the 2018-19 season, the BILs displayed delayed flowering and maturity. In the case of the OE2 trial, the mean days to heading increased by 9 days, and the

mean days to maturity increased by 11 days in comparison to the OE1 trial (Table 2). Similarly, in the heat-stressed trial HE2, the mean days to heading and days to maturity increased by 7 days in comparison to the HE1 trial (Table 2).

The degree of correlation among the various heat tolerance component traits under optimum and heat-stressed environments is shown schematically in Figure 3. The BLUPs of the phenotypic traits calculated over the two seasons were used to plot the correlation matrix. The plot yield showed a significant positive correlation with TGW, GPS, SN, PL and NDVI_H. A significant negative correlation was observed for plot yield with the DTH and DTM. The GFD displayed a negative and significant correlation with plot yield only under the optimum environment.

TABLE 3 Mean square values of the pooled analysis of variance for the optimum environment trial in the *T. durum*-*Ae. speltoides* BILs.

Source	df	PY	TGW	GPS	SN	SL	PL	DTH	DTM	GFD	NDVI_H
Genotype	227	210230.09**	13787.10**	165.61**	6.46**	1.57**	18.91**	36.02**	32.84**	16.66**	0.002**
Replication	1	57.75	2.55	15.47	11.79	1.48	0.63	1.04	8.42	3.53	0.04
Year	1	1904584.50**	47.57**	568.97**	75.05**	0.05	59.24**	17943.44**	32119.67**	2049.05**	0.03**
Genotype X Year	227	3765.14	539.66*	5.39**	1.03**	0.13*	0.65	2.01**	3.67**	1.13*	0.002*

PY, Plot yield; TGW, Thousand-grain weight; GPS, Grains per spike; SN, Spikelet number per spike; SL, Spike length; PL, Peduncle length; DTH, Days to heading; DTM, Days to maturity; GFD, Grain filling duration; NDVI_H, NDVI at heading.
**Significant at p<0.01, *Significant at p<0.05.

TABLE 4 Mean square values of the pooled analysis of variance for the heat-stressed environment trial in the *T. durum*-*Ae. speltoides* BILs.

Source	df	PY	TGW	GPS	SN	SL	PL	DTH	DTM	GFD	NDVI_H
Genotype	227	82762.81**	37.50**	132.10**	3.39**	1.25**	14.23**	17.11**	23.98**	16.64**	0.005**
Replication	1	203118.37	97.67	7.76	3.33	0.05	5.95	9.20	79.83	0.66	0.12
Year	1	1126440.07**	534.20**	566.97**	284.08**	1.38**	115.23**	9344.16**	11333.09**	95.88**	1.51**
Genotype X Year	227	6507.30**	0.85	6.41**	0.70**	0.10*	1.16**	1.42**	1.83**	34.83	0.004**

PY, Plot yield; TGW, Thousand-grain weight; GPS, Grains per spike; SN, Spikelet number per spike; SL, Spike length; PL, Peduncle length; DTH, Days to heading; DTM, Days to maturity; GFD, Grain filling duration; NDVI_H, NDVI at heading.
**Significant at $p < 0.01$, *Significant at $p < 0.05$.

3.2 Genotyping by sequencing and introgression profiling of the BILs

The distribution of the 2945 high-quality polymorphic SNPs along the 14 chromosomes is given in [Supplementary Figure S6](#). The maximum number of polymorphic SNP markers were mapped to chromosome 2B, while chromosome 1A had the least number of SNPs ([Table 5](#)). Introgression profiling of the BILs indicated multiple introgressions in both the A and B genomes of the durum wheat cv. PDW274 ([Supplementary Figure S7](#)). Introgressions were observed in all the chromosomes with variable fragment lengths. In some lines, we could observe the entire chromosomes of *Ae. speltoides* being substituted ([Supplementary Figure S7](#)). The percentage of *Ae. speltoides* introgressions in the 14 *T. durum* chromosomes ranged from 4.4 to 17.8%, with an overall mean of 10.5%.

3.3 QTL mapping for heat stress tolerance component traits

QTL mapping was carried out using the year-wise and over-the-year BLUPs of all the phenotypic traits except plant height for OE

and HE. In total, 30 QTLs were detected under individual and pooled environments. The list of QTLs detected for various heat tolerance component traits is presented in [Table 6](#). Out of these, 12 QTLs were detected under OE, 17 QTLs were detected under HE, and one QTL was detected in both OE and HE. The sign of the additive effect denotes the donor of QTLs, like the positive additive effect, which indicates the mapped SNP allele is derived from *Ae. speltoides*, while the negative additive effect represents the contribution by the *T. durum* parent. Concerning QTL density, chromosome 2B had the highest number of mapped QTLs (9 QTLs), and chromosome 1A, 3B, 4B and 6B had the least number (1 QTL each). The mapped QTLs along with their chromosomal positions are visually depicted in [Figure 4](#).

With respect to the OE, 12 QTLs were detected for SN, SL, PL, NDVI_H, TGW, GFD and DTM on chromosome 1B, 2A, 2B, 4B, 5B and 6A respectively ([Table 6](#)). Of these, 5 QTLs, *QSn.pau-As-OE-2B.2*, *QPl.pau-As-OE-2B.1*, *QTgw.pau-As-OE-5B*, *QGfd.pau-As-OE-6A* and *QDtm.pau-As-OE-1B* were consistently detected across the years ([Table 6](#)). The phenotypic variance explained (PVE) ranged from 5 to 11.2%. In the case of HE, 17 QTLs were detected for the SN, PY, NDVI_H, TGW, GPS, GFD, DTH and DTM on chromosomes 1A, 1B, 2A, 2B, 3B, 5A, 5B, 6A and 6B. QTLs that were detected across both the years include *QSn.pau-Td-*

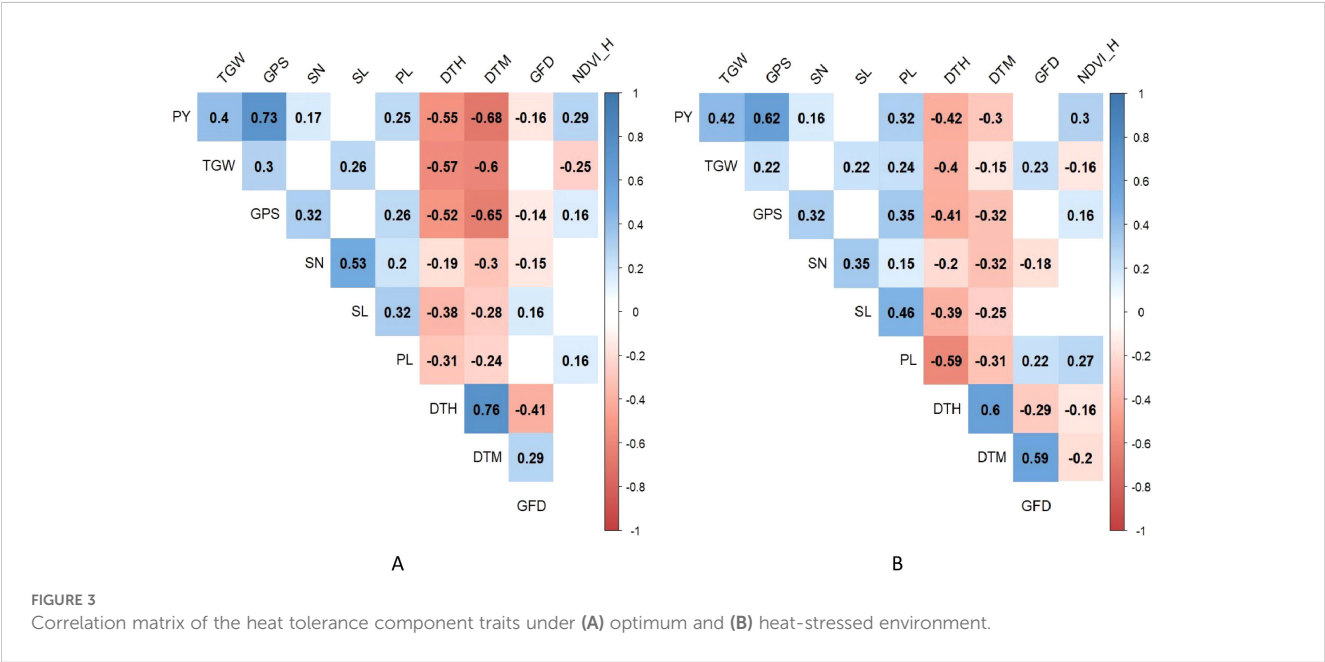


TABLE 5 Statistics of the distribution of the polymorphic SNPs along the 14 chromosomes of the *T. durum*-*Ae. speltoides* BILs.

Chromosome	Number of SNPs	% <i>T. durum</i> genome	% <i>Ae. speltoides</i> introgressions	% missing data
1A	114	86.4	8.5	5.1
1B	211	86.0	8.6	5.4
2A	217	78.9	16.2	4.9
2B	373	77.9	17.2	4.9
3A	114	83.0	12.5	4.5
3B	190	80.4	13.9	5.7
4A	127	88.2	7.0	4.8
4B	151	89.7	5.2	5.1
5A	258	81.5	13.3	5.2
5B	337	79.0	15.4	5.6
6A	136	88.2	7.1	4.7
6B	318	88.2	6.5	5.3
7A	149	88.8	5.6	5.6
7B	250	89.8	4.7	5.5
Total	2945	84.7	10.1	5.2

TABLE 6 List of QTLs detected for various heat tolerance component traits under the optimum environment (OE) and heat-stressed environment (HE) in the *T. durum*-*Ae. speltoides* BILs.

S. No.	QTL	Trait	SNP marker	Position (Mb)	Environment	LOD	PVE (%)	Additive effect
Optimum environment								
1	<i>QSn.pau-Td-OE-2B.1</i>	SN	2B_47840815	47.84	OE_2018-19	5.1	6.9	-0.55
					OE_pooled	3.6	5.0	-0.42
2	<i>QSn.pau-As-OE-2B.2</i>	SN	2B_752403466	752.40	OE_2017-18	4.2	6.3	0.36
					OE_2018-19	4.5	6.0	0.38
					OE_pooled	5.3	7.3	0.38
3	<i>QSn.pau-Td-OE-5B</i>	SN	5B_662962714	662.96	OE_2017-18	4.3	7.2	-0.57
					OE_pooled	3.4	5.2	-0.49
4	<i>QSl.pau-As-OE-1B</i>	SL	1B_392758436	392.76	OE_2017-18	4.5	8.4	0.17
5	<i>QPl.pau-As-OE-2A</i>	PL	2A_160196721	160.20	OE_pooled	21.9	6.2	1.59
6	<i>QPl.pau-As-OE-2B.1</i>	PL	2B_250066440	250.07	OE_2017-18	2.5	5.2	0.58
					OE_2018-19	12.8	6.6	1.27
					OE_pooled	18.0	5.7	1.52
7	<i>QPl.pau-Td-OE-2B.2</i>	PL	2B_649975448	649.97	OE_2018-19	22.2	11.2	-1.82
					OE_pooled	20.2	5.8	-1.71
8	<i>QNdvi-h.pau-Td-OE-4B</i>	NDVI_H	4B_80004193	80.00	OE_2018-19	10.7	8.3	-0.01
9	<i>QTgw.pau-As-OE-5B</i>	TGW	5B_350487145	350.49	OE_2017-18	3.1	7.5	1.30
					OE_2018-19	2.7	6.3	1.07
					OE_pooled	3.0	7.2	1.21

(Continued)

TABLE 6 Continued

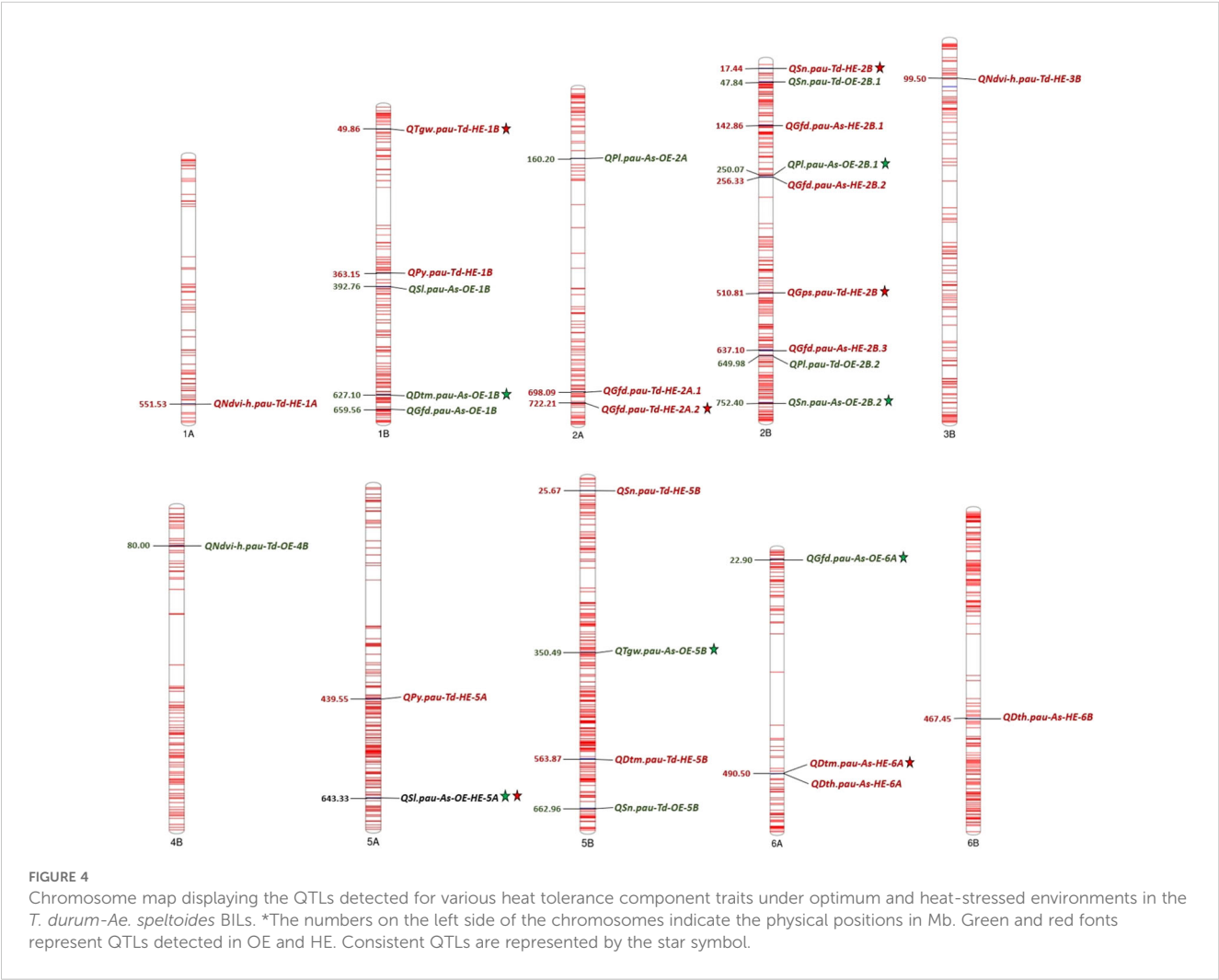
S. No.	QTL	Trait	SNP marker	Position (Mb)	Environment	LOD	PVE (%)	Additive effect
Optimum environment								
10	<i>QGfd.pau-As-OE-1B</i>	GFD	1B_659555349	659.55	OE_2017-18	4.2	5.4	0.93
11	<i>QGfd.pau-As-OE-6A</i>	GFD	6A_22900690	22.90	OE_2017-18	2.5	5.0	0.97
					OE_2018-19	5.0	5.1	1.26
					OE_pooled	3.9	5.3	1.01
12	<i>QDtm.pau-As-OE-1B</i>	DTM	1B_627098121	627.10	OE_2017-18	2.9	5.6	1.19
					OE_2018-19	7.8	7.5	1.86
					OE_pooled	11.4	8.4	2.27
Heat stressed environment								
13	<i>QSn.pau-Td-HE-2B</i>	SN	2B_17444376	17.44	HE_2017-18	2.5	5.1	-0.21
					HE_2018-19	3.0	5.1	-0.23
					HE_pooled	2.9	5.0	-0.24
14	<i>QSn.pau-Td-HE-5B</i>	SN	5B_25666540	25.67	HE_2017-18	3.3	6.2	-0.56
					HE_pooled	3.8	6.6	-0.50
15	<i>QPy.pau-Td-HE-1B</i>	PY	1B_363149235	363.15	HE_2018-19	7.9	8.7	-89.06
					HE_pooled	7.6	8.6	-88.84
16	<i>QPy.pau-Td-HE-5A</i>	PY	5A_439551042	439.55	HE_2017-18	2.9	6.1	-52.4
17	<i>QNdvi-h.pau-Td-HE-1A</i>	NDVI_H	1A_551533041	551.53	HE_2017-18	7.8	9.2	-0.01
18	<i>QNdvi-h.pau-Td-HE-3B</i>	NDVI_H	3B_99501367	99.50	HE_2018-19	9.3	6.7	-0.01
19	<i>QTgw.pau-Td-HE-1B</i>	TGW	1B_49863151	49.86	HE_2017-18	3.2	5.4	-1.46
					HE_2018-19	2.7	5.3	-1.27
					HE_pooled	2.8	5.5	-1.34
20	<i>QGps.pau-Td-HE-2B</i>	GPS	2B_510812119	510.81	HE_2017-18	3.2	6.1	-2.18
					HE_2018-19	2.5	5.0	-1.90
					HE_pooled	2.6	5.1	-1.97
21	<i>QGfd.pau-Td-HE-2A.1</i>	GFD	2A_698093562	698.09	HE_2017-18	8.1	6.2	-0.72
22	<i>QGfd.pau-Td-HE-2A.2</i>	GFD	2A_722208150	722.21	HE_2017-18	6.1	5.0	-0.67
					HE_2018-19	12.5	8.6	-1.11
					HE_pooled	8.5	5.3	-0.76
23	<i>QGfd.pau-As-HE-2B.1</i>	GFD	2B_142863752	142.86	HE_2018-19	13.3	8.3	1.12
24	<i>QGfd.pau-As-HE-2B.2</i>	GFD	2B_256328903	256.33	HE_2017-18	14.7	10.6	0.96
					HE_pooled	16.9	10.8	1.10
25	<i>QGfd.pau-Td-HE-2B.3</i>	GFD	2B_637100920	637.10	HE_pooled	9.4	6.0	-0.80
26	<i>QDtm.pau-Td-HE-5B</i>	DTM	5B_563875919	563.87	HE_2017-18	4.1	5.8	-0.96
					HE_pooled	3.4	5.0	-0.87
27	<i>QDtm.pau-As-HE-6A</i>	DTM	6A_490496402	490.50	HE_2017-18	4.4	6.4	0.78
					HE_2018-19	3.8	5.0	0.72
					HE_pooled	4.3	6.1	0.74

(Continued)

TABLE 6 Continued

S. No.	QTL	Trait	SNP marker	Position (Mb)	Environment	LOD	PVE (%)	Additive effect
Heat stressed environment								
28	<i>QDth.pau-As-HE-6A</i>	DTH	6A_490496402	490.50	HE_2017-18	2.9	7.2	0.54
29	<i>QDth.pau-As-HE-6B</i>	DTH	6B_467448196	467.45	HE_2018-19	9.0	6.3	2.11
					HE_pooled	9.7	6.7	2.04
Optimum and Heat stressed environment								
30	<i>QSl.pau-As-OE-HE-5A</i>	SL	5A_643325049	643.32	OE_2018-19	3.6	8.4	0.27
					OE_pooled	3.6	8.3	0.25
					HE_2017-18	3.4	6.3	0.22
					HE_2018-19	4.9	9.6	0.26
					HE_pooled	4.4	8.5	0.24

PY, Plot yield; TGW, Thousand-grain weight; GPS, Grains per spike; SN, Spikelet number per spike; SL, Spike length; PL, Peduncle length; DTH, Days to heading; DTM, Days to maturity; GFD, Grain filling duration; NDVI_H, NDVI at heading.



HE-2B, *QTgw.pau-Td-HE-1B*, *QGps.pau-Td-HE-2B*, *QGfd.pau-Td-HE-2A.2* and *QDtm.pau-As-HE-6A* (Table 6). The PVE of the detected QTLs ranged from 5 to 10.8%. The QTL for SL *QSl.pau-As-OE-HE-5A* on chromosome 5A was detected under both OE and HE, with the PVE ranging from 6.3 to 9.6%.

3.4 Testing the allelic effect of the SNP associated with the stable QTLs

The effect contributed by the alleles linked to the stable QTLs was studied by comparing the phenotypic distributions of the genotype groups sorted based on the allelic pattern at the SNP locus. The stable QTLs detected under the HE were used for the analysis. The Mann-Whitney U test revealed significant differences in the median phenotypic performance among the alternative allelic groups for four out of the five stable QTLs tested (Figure 5). The difference in the median phenotypic performance of the allelic group was insignificant in the case of the QTL for GFD *QGfd.pau-Td-HE-2A.2*. The SNP allele linked to all these stable QTLs, excluding *QDtm.pau-As-HE-6A*, was contributed by *T. durum* parent.

3.5 Postulation of putative candidate genes in the significant QTL regions mapped under the heat-stressed environment

The genome-wide LD decay value was estimated to be 1.85 Mb (Supplementary Figure S8). The 1.85 Mb physical region flanking the SNPs linked to the consistent QTLs detected under the HE revealed the presence of multiple genes known to modulate heat stress responses. The annotations of the postulated candidate genes are given in Table 7. A total of 21 genes were selected after scanning the 1.85 Mb interval flanking the five consistent QTLs based on the gene annotation and literature search. The proteins encoded by these putative candidate genes include cytochrome P450, lectin receptor kinase, E3 ubiquitin-protein ligase, arginine/serine-rich splicing factor, ABC transporter protein, peptidyl-prolyl cis-trans isomerase, alcohol dehydrogenase, F-box protein, plastid-lipid associated protein and MYB family protein.

4 Discussion

4.1 Phenotypic evaluation of BILs for heat stress tolerance and introgression profiling

The BILs displayed significant variation for all the heat tolerance component traits studied. Due to the unseasonal rainfall during the 2018-19 trial, the HE2 trial recorded a decline in the plot yield and NDVI. This was due to the reduction in viability and growth of the BILs sown in the HE2 trial. Correlation analysis indicated that the early flowering and early maturing genotypes produce better yields during terminal heat stress possibly due to

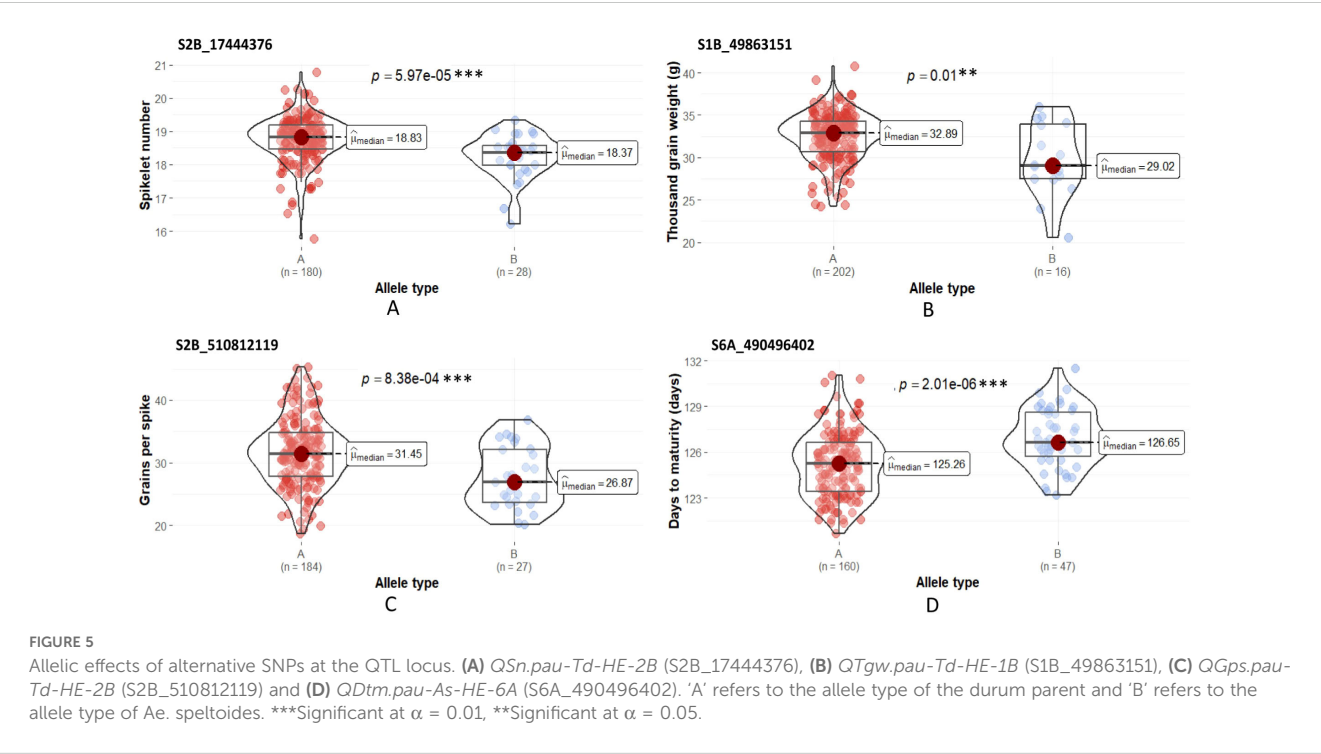
escape from high temperature stress. Selection for higher TGW, GPS, SN, PL and NDVI_H can help in achieving better yields in both environments.

Introgression profiling of the BILs revealed random introgressions of donor chromosome fragments across all the 14 chromosomes of the recurrent parent. This phenomenon was observed likely due to the *Ph1* suppression activity of the *Ae. speltooides* accession pau3809, which would have induced the crossing over between the S genome of *Ae. speltooides* and both the A and B genomes of the durum wheat cv. PDW274 (Dvorak et al., 2006; Millet, 2008).

4.2 QTL mapping

The QTLs were mapped under OE and HE for the two cropping seasons. Out of the 12 QTLs detected under OE, five QTLs were detected consistently over the two seasons. The consistent QTL for SN *QSn.pau-As-OE-2B.2*, contributed by *Ae. speltooides*, was mapped on chromosome 2B. Reports on the mapping of SN QTLs on chromosome 2B are available in the studies of Edae et al. (2014) and Li et al. (2021). Stable QTL for PL *QPl.pau-As-OE-2B.1* was mapped on chromosome 2B, as reported earlier by Neumann et al., 2011 and Graziani et al., 2014. A stable QTL for TGW *QTgw.pau-As-OE-5B* was located on chromosome 5B at a physical distance of 350.49 Mb, with a maximum PVE of 7.5%. A major and stable QTL for TGW *QTgw.caas-5B* was fine-mapped in the 5B chromosome within the physical interval of 49.6 Mb-51.6 Mb in a RIL population derived from Zhongmai871 and Zhongmai895 (Zhao et al., 2021). Results on mapping TGW QTLs on the same chromosome were obtained by Ramya et al. (2010); Zhang et al. (2018); Dhakal et al. (2021); Amalova et al. (2021) and Ahmed et al. (2022). The stable QTL controlling grain filling duration was detected on the 6A chromosome and localized to a physical interval of 22.90 Mb. To the best of our knowledge, this is the first report of mapping a GFD QTL on the 6A chromosome. A meta-QTL MQTL6A.3 for multiple traits, including grain number, spike-related traits, grain morphology-related traits and days to heading, was found to span this QTL and was delimited within the physical interval of 12.4 Mb to 43.4 Mb (Saini et al., 2022). The stable QTL for DTM *QDtm.pau-As-OE-1B* was detected on chromosome 1B, with the PVE ranging from 5.6 to 8.4%. Similar results are available in the studies of Cuthbert et al. (2008) and Kamran et al. (2013).

In the case of HE, 17 QTLs were detected in total, and among those, five QTLs were detected across the seasons. Consistent QTLs were detected for the traits SN, TGW, GPS, GFD and DTM. The stable QTL for SN *QSn.pau-Td-HE-2B* was detected on chromosome 2B. Earlier workers had reported mapping of SN QTLs on chromosome 2B (Qaseem et al., 2019; Telfer et al., 2021). The stable TGW QTL *QTgw.pau-Gw-Td-HE-1B* was detected on chromosome 1B at a physical distance of 49.86 Mb with a maximum PVE of 5.5%. Pankaj et al. (2024) identified a stable QTL for TGW located close to the mapped QTL and flanked by the markers *Xwmc44* and *Xwmc367* on chromosome 1B. Tahmasebi et al. (2016) and Ogbonnaya et al. (2017) also detected TGW QTLs



on the same chromosome. The consistent QTL for GPS *QGps.pau-Td-HE-2B* was mapped to the 2B chromosome, with the PVE ranging from 5 to 6.1%. Previous reports are available on mapping of GPS QTLs on chromosome 2B (Tahmasebi et al., 2016; Sharma et al., 2016; Li et al., 2019; Elbashier et al., 2023). The consistent QTL for GFD *QGfd.pau-Td-HE-2A.2* was located on chromosome 2A at a physical distance of 722.21 Mb with a maximum PVE of 12.5%. Reports on the mapping of GFD QTLs on chromosome 2A are found in the publications of Mason et al. (2010) and Pankaj et al. (2024). The photoperiod sensitivity gene *Ppd-A1*, responsible for the

TABLE 7 List of postulated candidate genes for the consistent QTLs detected under the heat-stressed environment in the *T. durum*-Ae. speltoides BILs.

S. No.	QTL	Trait	Chromosome	Mapped SNP position (Mb)	Distance from SNP (Kb)	Gene stable ID	Function
1	<i>QSn.pau-Td-HE-2B</i>	SN	2B	17.44	-333.773	<i>TraesCS2B01G035600</i>	Cytochrome P450
					-300.309	<i>TraesCS2B01G035700</i>	Cytochrome P450
					+55.28	<i>TraesCS2B01G036600</i>	Lectin receptor kinase
					-4.197	<i>TraesCS2B01G036900</i>	Cytochrome P450
					-0.51	<i>TraesCS2B01G037000</i>	Cytochrome P450
					+58.886	<i>TraesCS2B01G037600</i>	Cytochrome P450
					+70.952	<i>TraesCS2B01G037700</i>	Cytochrome P450
2	<i>QTgw.pau-Td-HE-1B</i>	TGW	1B	49.86	-643.722	<i>TraesCS1B01G064800</i>	Cytochrome P450
					-69.26	<i>TraesCS1B01G065800</i>	E3 ubiquitin-protein ligase RNF14
					-320.63	<i>TraesCS1B01G065900</i>	Arginine/serine-rich splicing factor
					-340.66	<i>TraesCS1B01G066000</i>	ABC transporter G family member

(Continued)

TABLE 7 Continued

S. No.	QTL	Trait	Chromosome	Mapped SNP position (Mb)	Distance from SNP (Kb)	Gene stable ID	Function
					+537.015	<i>TraesCS1B01G066300</i>	Peptidyl-prolyl cis-trans isomerase
					+911.743	<i>TraesCS1B01G066400</i>	Alcohol dehydrogenase
3	<i>QGps.pau-Td-HE-2B</i>	GPS	2B	510.81	-754.56	<i>TraesCS2B01G357000</i>	F-box family protein
					-320.48	<i>TraesCS2B01G358200</i>	ABC transporter ATP-binding protein
					+418.56	<i>TraesCS2B01G357200</i>	Cytochrome P450
4	<i>QGfd.pau-Td-HE-2A.2</i>	GFD	2A	722.21	+0.096	<i>TraesCS2A01G487600</i>	F-box protein
					-311.44	<i>TraesCS2A01G487900</i>	Plastid-lipid associated protein/fibrillin family protein
					+516.360	<i>TraesCS2A01G488200</i>	MYB-like transcription factor family protein
					+894.275	<i>TraesCS2A01G489200</i>	F-box protein
5	<i>QDtm.pau-As-HE-6A</i>	DTM	6A	490.49	+194.61	<i>TraesCS6A01G265300</i>	MYB family protein

(Positive sign of the distance from SNP denotes that the gene was present upstream of the mapped QTL (SNP) position and negative sign indicates that the gene was present downstream of the mapped QTL (SNP) position).

TGW, Thousand-grain weight; GPS, Grains per spike; SN, Spikelet number per spike; DTM, Days to maturity; GFD, Grain filling duration.

initiation of flowering in wheat, was reported to be located on the short arm of chromosome 2A, and mutations in this gene enable the plants to flower early (Law et al., 1978; Wilhelm et al., 2009; Achilli et al., 2022). Early flowering spring wheat genotypes tend to have longer grain filling durations and produce better yields, especially under terminal heat stress (Vinod et al., 2012). For the trait DTM, a stable QTL was detected on chromosome 6A and a similar result was documented by Qaseem et al. (2019).

A QTL for SL *QSL.pau-As-OE-HE-5A* was detected consistently across the years under both OE and HE. Many reports are available on mapping of SL QTL on chromosome 5A under non-stress conditions (Kumar et al., 2007; Wang et al., 2011; Ma et al., 2014; Kang et al., 2020). A stable QTL for SL was localized to a 6.69 Mb interval (518.4–525.1 Mb) on chromosome 5A, explaining 7.8 to 26.6% PVE (Ji et al., 2021). Under heat-stressed conditions, no published studies are available on detecting SL QTLs on this chromosome, suggesting it may be a novel QTL.

4.3 Allelic effect of the SNP associated with the stable QTLs on the linked phenotypic trait

The effect of the SNP allele associated with the stable QTLs mapped under HE on the phenotypic performance of the BILs was compared with the alternative allele using Mann-Whitney U test. The allelic effects were significant for four out of the five stable QTLs. These results validate that the QTL alleles significantly contributed to the enhancement of the associated phenotypic trait.

4.4 Postulation of putative candidate genes in the significant QTL regions mapped under the heat-stressed environment and their role in heat stress

Several genes were found in the 1.85 Mb physical interval flanking the five consistent QTLs detected under the HE. The gene annotations were used to perform a literature search to identify their role in heat stress regulation. Accordingly, 21 putative candidate genes were shortlisted and their role in mediating heat stress response will be discussed in the following sections.

Seven genes were found in the physical interval flanking the QTL for SN *QSn.pau-Td-HE-2B*, encoding cytochrome P450 and lectin receptor kinase. Shumayla et al. (2016) conducted a global expression analysis of bread wheat lectin receptor kinases (LRK) genes and identified 263 LRKs, out of which 77 LRKs were differentially expressed under heat, drought or a combination of both stresses. About 40% of the rice Lectin RLKs identified from the genome-wide analysis were differentially expressed during heat stress treatment (Vaid et al., 2012). Eighteen Lectin RLK genes were upregulated in foxtail millet plants treated with 6% polyethylene glycol (PEG-6000), simulating drought and high temperature stress conditions (Zhao et al., 2016). Multiple transcripts of Cytochrome P450 (CYP) were up- or down-regulated during the heat stress treatment at 40 °C for 10 hour in perennial ryegrass and tall fescue (Tao et al., 2017). Heat stress induced the expression of 11 different CYP genes in switchgrass, and two genes belonging to the CYP71A1 family involved in the secretion of indole alkaloid secologanin were upregulated (Li et al., 2013a). The hormone abscisic acid (ABA)

plays a role in the arrest of plant growth during biotic or abiotic stresses and growth phase transitions in plant development (Umezawa et al., 2010). Genes responsible for ABA catabolism (*CYP707A* family genes) were downregulated during heat stress in Arabidopsis (Baron et al., 2012).

The region flanking the stable QTL *QTgw.pau-Td-HE-1B* was found to contain 6 genes coding for cytochrome P450, ubiquitin-protein ligase RNF14, arginine/serine-rich splicing factor, ABC transporter G family member, peptidyl-prolyl cis-trans isomerase and alcohol dehydrogenase. The rice *OsDHSRP1* gene, a RING finger E3 ligase, negatively regulates heat, drought and salt stress responses by degrading the glyoxalase protein through the Ub/26S proteasome system (Kim et al., 2020). Liu et al. (2020) reported the involvement of Arabidopsis *AtPPRT1* gene, a zinc finger ubiquitin E3 ligase, in conferring basal and acquired thermotolerance by mediating the degradation of uncharacterized proteins. T-DNA insertion mutants for the *AtPPRT1* gene displayed lower germination and survival rates when exposed to high-temperature stress. Palusa and Reddy (2015) elucidated the involvement of spliced isoforms of serine- and arginine-rich (SR) proteins during heat stress in Arabidopsis. Heat priming in Arabidopsis induced the expression of alternatively spliced heat shock factors (HSFs) and SR transcripts *SR30*, *SR45a*, *SR34*, and *RS41*, implicating the role of these proteins in stress memory (Ling et al., 2018). Exposure of Arabidopsis plants to combined drought and heat stress upregulated the expression levels of the ATP-binding cassette (ABC) transporter gene *At1g64550* (Rizhsky et al., 2004). Gupta et al. (2019) characterized the plant-specific ABC receptors termed ABCG receptors in rice and identified multiple ABC receptor genes showing stress-specific expression patterns under various abiotic stresses like drought, heat and salinity. Wang et al. (2016) imposed heat priming on wheat plants at seedling and flowering stages and observed an upregulation of the *cyclophilin 38* gene, a type of peptidyl-prolyl cis-trans isomerase (PPI) when the progeny of primed plants were exposed to post anthesis heat stress. Genome-wide analysis of cyclophilin genes in wheat led to the identification of 83 cyclophilin genes distributed in all 21 wheat chromosomes. Some of these genes housed a heat shock element in their regulatory region and were known to modulate heat stress responses (Singh et al., 2019). Li et al. (2013b) carried out a proteomic analysis of the leaf tissues of alfalfa plants exposed to different periods of heat stress. The alcohol dehydrogenase (ALDH) protein displayed a four-fold higher expression under a 48-hour heat stress treatment compared to the control plants. The Arabidopsis ALDH genes *ALDH3I1* and *ALDH7B4*, known to participate in various abiotic stress responses, were strongly induced by high-temperature stress, and the double knockout mutants displayed severe heat sensitivity (Zhao et al., 2017a).

The QTL for GPS *QGps.pau-Td-HE-2B* houses three genes encoding the F-box family protein, ABC transporter ATP-binding protein and cytochrome P450. The functions of ABC transporter protein and cytochrome P450 are already discussed in the preceding sections. Stefanowicz et al. (2015) elaborated the role of plant F-box proteins as regulatory players of proteolytic mechanisms in response to cellular stimuli during hormone signalling, abiotic and biotic stresses and plant morphogenesis. Li et al. (2018) overexpressed the wheat F-box gene *TaFBA1* in transgenic

tobacco plants, followed by exposure of the transgenics to heat stress. The transgenic plants exhibited lesser growth inhibition and increased photosynthesis compared to the control. Genes involved in ROS scavenging, proline biosynthesis and abiotic stress regulation were upregulated in the transgenic plants.

The grain filling duration QTL *QGfd.pau-Td-HE-2A.2* encompasses four genes coding for F-box protein, plastid-lipid associated protein/fibrillin family protein and MYB-like transcription factor family protein, respectively, within the 1.85 Mb interval. Genome-wide analysis and expression characterization of rice fibrillin genes (*OsFBNs*), led to the identification of 11 fibrillin genes (*OsFBN1* to *OsFBN11*) that were upregulated during heat stress (Li et al., 2020). Cis-acting elements involved in hormonal signalling, photoreaction and environmental stresses were detected upstream to these genes. Jiang et al. (2020) identified and characterized 26 FBN genes distributed on 11 wheat chromosomes using genome-wide analysis and expression studies. The wheat FBN genes *TaFBN-A1*, *TaFBN-B1*, *TaFBN-A2*, *TaFBN-B2*, *TaFBN-D2*, *TaFBN-D6* and *TaFBN-B6* were significantly upregulated during abiotic and biotic stress treatments, including heat, implicating their role in stress responses. MYB30, an R2R3 type MYB transcription factor, regulates oxidative and heat stress responses in Arabidopsis by controlling cytosolic calcium signalling (Liao et al., 2017). The mutant plants for the *MYB30* gene were sensitive to heat stress. Six heat-induced MYB genes (*TaMYBs*) were identified from transcriptome data derived from wheat plants subjected to heat stress (Zhao et al., 2017b). One of the MYB genes, *TaMYB80*, conferred heat and drought tolerance when expressed in Arabidopsis on account of increased cellular levels of ascorbic acid. Jacob et al. (2021) identified two MYB transcription factor genes, *TT2* and *MYB5*, that are co-regulated along with Heat Shock Factor A2, which is a regulator of multiple environmental stresses in Arabidopsis. Overexpression of the *TT2* and *MYB5* genes conferred enhanced heat stress tolerance in the transgenic plants, whereas the *tt2*, *myb5* and *tt2/myb5* loss of function mutants displayed heat sensitivity. A gene encoding MYB family protein was located in the 1.85 Mb interval flanking the stable QTL for DTM *QDtm.pau-As-HE-6A*. The role of this gene was discussed in the previous sections.

5 Conclusion

Negative effects of heat stress on wheat yields has caused serious concerns for wheat growers in recent times. For wheat production to be sustained in the near future, it is essential to identify genetic regions contributing to heat tolerance in the exotic and wild wheat germplasm. In an attempt to uncover genomic regions controlling heat tolerance-related traits, we conducted a QTL mapping study on a set of *T. durum*-*Ae. speltoides* BILs. Thirty QTLs were detected for various heat tolerance component traits under optimum and heat-stressed environments in multiple wheat chromosomes. Stable QTLs were identified for some of the traits under heat-stressed environment for use in marker-assisted selection (MAS). The 1.85 Mb physical interval flanking the stable heat tolerance QTLs harboured genes known to play essential roles in regulating heat stress responses. The BILs carrying multiple heat tolerance QTL can

be employed in wheat breeding programmes for improving heat tolerance using MAS.

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.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any identifiable images or data included in this article.

Author contributions

NK: Investigation, Methodology, Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Software. SK: Formal analysis, Investigation, Methodology, Writing – review & editing, Resources, Supervision, Validation. UK: Formal analysis, Investigation, Methodology, Supervision, Writing – review & editing, Data curation. RS: Investigation, Methodology, Writing – review & editing, Validation. GD: Investigation, Methodology, Writing – review & editing, Data curation, Formal analysis, Software. PB: Data curation, Formal analysis, Methodology, Resources, Writing – review & editing. PC: Methodology, Writing – review & editing, Conceptualization, Funding acquisition, Investigation, Resources, Supervision, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research related to this article. The financial support (Grant no. BT/

01/CEIB/12/I/01) provided by the Department of Biotechnology, Ministry of Science and Technology, Government of India in the form of DBT Programme Support is gratefully acknowledged.

Acknowledgments

The authors immensely thank Dr. Jesse Poland, Professor, Plant Science, King Abdullah University of Science and Technology, Thuwal, Kingdom of Saudi Arabia for his help in the genotyping of the mapping population. The first author NK acknowledges the financial assistance received from the Department of Biotechnology, Government of India in the form of a Junior Research Fellowship and Senior Research Fellowship.

Conflict of interest

Author UK was employed by Astralyn Agro OPC Pvt. Ltd.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2024.1485914/full#supplementary-material>

References

- Achilli, A. L., Roncallo, P. F., Larsen, A. O., Dreisigacker, S., and Echenique, V. (2022). Population structure, allelic variation at Rht-B1 and Ppd-A1 loci and its effects on agronomic traits in Argentinian durum wheat. *Sci. Rep.* 12, 9629. doi: 10.1038/s41598-022-13563-w
- Adamski, N. M., Borrill, P., Brinton, J., Harrington, S. A., Marchal, C., Bentley, A. R., et al. (2020). A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. *Elife* 9, e55646. doi: 10.7554/eLife.55646
- Ahmed, H. G. M. D., Naeem, M., Zeng, Y., Rashid, M. A. R., Ullah, A., Saeed, A., et al. (2022). Genome-wide association mapping for high temperature tolerance in wheat through 90k SNP array using physiological and yield traits. *PLoS One* 17, e0262569. doi: 10.1371/journal.pone.0262569
- Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S., and Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* 1, 2320–2325. doi: 10.1038/nprot.2006.384
- 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, 745–756. doi: 10.1016/j.cj.2020.03.010
- Amalova, A., Abugalieva, S., Chudinov, V., Sereda, G., Tokhetova, L., Abdikhalyk, A., et al. (2021). QTL mapping of agronomic traits in wheat using the UK Avalon × Cadenza reference mapping population grown in Kazakhstan. *PeerJ* 9, e10733. doi: 10.7717/peerj.10733
- Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., Rogers, J., et al. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* 361, 6403. doi: 10.1126/science.aar7191
- Asseng, S., Ewert, F., Martre, P., Rötter, R. P., Lobell, D. B., Cammarano, D., et al. (2015). Rising temperatures reduce global wheat production. *Nat. Clim. Change* 5, 143–147. doi: 10.1038/nclimate2470
- Asithir, B. (2015). Mechanisms of heat tolerance in crop plants. *Biol. Plant* 59, 620–628. doi: 10.1007/s10535-015-0539-5

- Awlchew, Z. T., Singh, R., Kaur, S., Bains, N. S., and Chhuneja, P. (2016). Transfer and mapping of the heat tolerance component traits of *Aegilops speltoides* in tetraploid wheat *Triticum durum*. *Mol. Breed.* 36, 78. doi: 10.1007/s11032-016-0499-2
- Balla, K., Karsai, I., Bencze, S., and Veisz, O. (2012). Germination ability and seedling vigour in the progeny of heat-stressed wheat plants. *J. Acta Agron. Hung.* 60, 299–308. doi: 10.1556/AAgr.60.2012.4.1
- Baron, K. N., Schroeder, D. F., and Stasolla, C. (2012). Transcriptional response of abscisic acid (ABA) metabolism and transport to cold and heat stress applied at the reproductive stage of development in *Arabidopsis thaliana*. *Plant Sci.* 188, 48–59. doi: 10.1016/j.plantsci.2012.03.001
- Bhullar, R., Nagarajan, R., Bennypaul, H., Sidhu, G. K., Sidhu, G., Rustgi, S., et al. (2014). Silencing of a metaphase I-specific gene results in a phenotype similar to that of the Pairing homeologous 1 (Ph1) gene mutations. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14187–14192. doi: 10.1073/pnas.1416241111
- Bhusal, N., Sarial, A. K., Sharma, P., and Sareen, S. (2017). Mapping QTLs for grain yield components in wheat under heat stress. *PLoS One* 12, e0189594. doi: 10.1371/journal.pone.0189594
- Boden, S. A., McIntosh, R. A., Uauy, C., Krattinger, S. G., Dubcovsky, J., Rogers, W. J., et al. (2023). Updated guidelines for gene nomenclature in wheat. *Theor. Appl. Genet.* 136, 72. doi: 10.1007/s00122-023-04253-w
- Crespo-Herrera, L. A., Singh, R. P., Sabraoui, A., and El-Bouhssini, M. (2019). Resistance to insect pests in wheat—rye and *Aegilops speltoides* Tausch translocation and substitution lines. *Euphytica* 215, 1–7. doi: 10.1007/s10681-019-2449-7
- Curtis, B. C., Rajaram, S., and Gómez, M. (2002). *Bread wheat: improvement and production* (Rome, Italy: Food and Agriculture Organization of the United Nations (FAO)).
- Cuthbert, J. L., Somers, D. J., Brûlé-Babel, A. L., Brown, P. D., and Crow, G. H. (2008). Molecular mapping of quantitative trait loci for yield and yield components in spring wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 117, 595–608. doi: 10.1007/s00122-008-0804-5
- Dhakal, S., Liu, X., Chu, C., Yang, Y., Rudd, J. C., Ibrahim, A. M., et al. (2021). Genome-wide QTL mapping of yield and agronomic traits in two widely adapted winter wheat cultivars from multiple mega-environments. *PeerJ* 9, e12350. doi: 10.7717/peerj.12350
- Dvorak, J., Deal, K. R., and Luo, M. C. (2006). Discovery and mapping of wheat Ph1 suppressors. *Genetics* 174, 17–27. doi: 10.1534/genetics.106.058115
- Eade, E. A., Byrne, P. F., Haley, S. D., Lopes, M. S., and Reynolds, M. P. (2014). Genome-wide association mapping of yield and yield components of spring wheat under contrasting moisture regimes. *Theor. Appl. Genet.* 127, 791–807. doi: 10.1007/s00122-013-2257-8
- Ehdaie, B., and Waines, J. G. (1992). Heat resistance in wild *Triticum* and *Aegilops*. *J. Genet. Breed.* 46, 221–228.
- Elbasher, E. M., Idris, S. E., Tahir, I. S., Elhanafi, S., Saad, A. S. I., Mustafa, H. M., et al. (2023). Genome wide association study of yield and yield-related traits in elite spring bread wheat genotypes grown under high temperature environment in Sudan. *Am. J. Plant Sci.* 14, 202–219. doi: 10.4236/ajps.2023.142016
- Farhad, M., Kumar, U., Tomar, V., Bhati, P. K., Krishnan J. N., Barek, V., et al. (2023). Heat stress in wheat: a global challenge to feed billions in the current era of the changing climate. *Front. Sustain. Food Syst.* 7. doi: 10.3389/fsufs.2023.1203721
- Gill, B. S., Friebe, B., Raupp, W. J., Wilson, D. L., Cox, T. S., Sears, R. G., et al. (2006). Wheat genetics resource center: the first 25 years. *Adv. Agron.* 89, 73–136. doi: 10.1016/S0065-2113(05)89002-9
- Glaubit, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: a high-capacity genotyping by sequencing analysis pipeline. *PLoS One* 9, e90346. doi: 10.1371/journal.pone.0090346
- Government of India (GOI) (2023). *The Final Estimates of production of major crops for the year 2022-23* (Government of India, New Delhi: Economics and Statistics Division, Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare).
- Graziani, M., Maccaferri, M., Royo, C., Salvatorelli, F., and Tuberosa, R. (2014). QTL dissection of yield components and morpho-physiological traits in a durum wheat elite population tested in contrasting thermo-pluviometric conditions. *Crop Pasture Sci.* 65, 80–95. doi: 10.1071/CP13349
- Gupta, B. B., Selter, L. L., Baranwal, V. K., Arora, D., Mishra, S. K., Sirohi, P., et al. (2019). Updated inventory, evolutionary and expression analyses of G (PDR) type ABC transporter genes of rice. *Plant Physiol. Biochem.* 142, 429–439. doi: 10.1016/j.plaphy.2019.08.004
- Hammer, Ø., Harper, D. A., and Ryan, P. D. (2001). PAST: Paleontological statistics software package for education and data analysis. *Palaeontol. Electronica* 4, 9.
- Hays, D., Mason, E., Do, J. H., Menz, M., and Reynolds, M. (2007). “Expression quantitative trait loci mapping heat tolerance during reproductive development in wheat (*Triticum aestivum*),” in *Wheat production in stressed environments*. Eds. H. T. Buck, J. E. Nisi and N. Salomón (Springer, Dordrecht), 373–382. doi: 10.1007/1-4020-5497-1_46
- Iqbal, M., Raja, N. I., Yasmeen, F., Hussain, M., Ejaz, M., and Shah, M. A. (2017). Impacts of heat stress on wheat: a critical review. *Adv. Crop Sci. Technol.* 5, 01–09. doi: 10.4172/2329-8863.1000251
- Jacob, P., Brisou, G., Dalmais, M., Thévenin, J., van der Wal, F., Latrasse, D., et al. (2021). The seed development factors TT2 and MYB5 regulate heat stress response in *Arabidopsis*. *Genes* 12, 746. doi: 10.3390/genes12050746
- Ji, G., Xu, Z., Fan, X., Zhou, Q., Yu, Q., Liu, X., et al. (2021). Identification of a major and stable QTL on chromosome 5A confers spike length in wheat (*Triticum aestivum* L.). *Mol. Breed.* 41, 1–13. doi: 10.1007/s11032-021-01249-6
- Jiang, Y., Hu, H., Ma, Y., and Zhou, J. (2020). Genome-wide identification and characterization of the fibrillin gene family in *Triticum aestivum*. *PeerJ* 8, e9225. doi: 10.7717/peerj.9225
- Joshi, N. L., and Kar, A. (2009). Contingency crop planning for dryland areas in relation to climate change. *Indian J. Agron.* 54, 237–243. doi: 10.59797/ija.v54i2.4787
- Kamran, A., Iqbal, M., Navabi, A., Randhawa, H., Pozniak, C., and Spaner, D. (2013). Earliness per se QTLs and their interaction with the photoperiod insensitive allele Ppd-D1a in the Cutler × AC Barrie spring wheat population. *Theor. Appl. Genet.* 126, 1995–1976. doi: 10.1007/s00122-013-2110-0
- Kang, C. S., Mo, Y. J., Kim, K. M., Kim, K. H., Chun, J. B., Park, C. S., et al. (2020). Mapping quantitative trait loci for yield potential traits in wheat recombinant inbred lines. *Agronomy* 11, 22. doi: 10.3390/agronomy11010022
- Khanna-Chopra, R., and Viswanathan, C. (1999). Evaluation of heat stress tolerance in irrigated environment of T. aestivum and related species. I. Stability in yield and yield components. *Euphytica* 106, 169–180. doi: 10.1023/A:1003531722420
- Khatkar, B. S., Chaudhary, N., and Dang, P. (2016). “Production and consumption of grains: India,” in *Encyclopedia of food grains, 2nd edition*. Eds. C. Wrigley, H. Corke, K. Seetharaman and J. Faubion (Academic Press, Oxford), 367–373. doi: 10.1016/B978-0-08-100596-5.00044-5
- Kim, S. K., Kim, J. H., and Jang, W. C. (2017). “Past, present and future molecular approaches to improve yield in wheat,” in *Wheat Improvement, Management and Utilization*. Eds. R. Wanyera and J. Owuochi (London IntechOpen). doi: 10.5772/67112
- Kim, J. H., Lim, S. D., and Jang, C. S. (2020). *Oryza sativa* drought-, heat- and salt-induced RING finger protein 1 (OsDHSRP1) negatively regulates abiotic stress-responsive gene expression. *Plant Mol. Biol.* 103, 235–252. doi: 10.1007/s11103-020-00989-x
- Kishii, M. (2019). An update of recent use of *Aegilops* species in wheat breeding. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00585
- Krishnan, J. N., Kumar, U., Kaur, S., and Chhuneja, P. (2023). Development of a novel image-based grain counting setup for thousand-grain weight estimation in wheat. *Indian J. Genet. Plant Breed.* 83, 469–475. doi: 10.31742/ISGPB.83.4.2
- Kumar, N., Kulwal, P. L., Balyan, H. S., and Gupta, P. K. (2007). QTL mapping for yield and yield contributing traits in two mapping populations of bread wheat. *Mol. Breed.* 19, 163–177. doi: 10.1007/s11032-006-9056-8
- Law, C., Snape, J., and Worland, A. (1978). The genetical relationship between height and yield in wheat. *Heredity* 40, 133–151. doi: 10.1038/hdy.1978.13
- Li, T., Deng, G., Tang, Y., Su, Y., Wang, J., Cheng, J., et al. (2021). Identification and validation of a novel locus controlling spikelet number in bread wheat (*Triticum aestivum* L.). *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.611106
- Li, J., Li, X., Khatib, A. A., and Xie, G. (2020). Phylogeny, structural diversity and genome-wide expression analysis of fibrillin family genes in rice. *Phytochemistry* 175, 112377. doi: 10.1016/j.phytochem.2020.112377
- Li, L., Mao, X., Wang, J., Chang, X., Reynolds, M., and Jing, R. (2019). Genetic dissection of drought and heat-responsive agronomic traits in wheat. *Plant Cell. Environ.* 42, 2540–2553. doi: 10.1111/pce.13577
- Li, Y. F., Wang, Y., Tang, Y., Kakani, V. G., and Mahalingam, R. (2013a). Transcriptome analysis of heat stress response in switchgrass (*Panicum virgatum* L.). *BMC Plant Biol.* 13, 1–12. doi: 10.1186/1471-2229-13-153
- Li, Q., Wang, W., Wang, W., Zhang, G., Liu, Y., Wang, Y., et al. (2018). Wheat F-box protein gene TaFBA1 is involved in plant tolerance to heat stress. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00521
- Li, W., Wei, Z., Qiao, Z., Wu, Z., Cheng, L., and Wang, Y. (2013b). Proteomics analysis of alfalfa response to heat stress. *PLoS One* 8, e82725. doi: 10.1371/journal.pone.0082725
- Liao, C., Zheng, Y., and Guo, Y. (2017). MYB30 transcription factor regulates oxidative and heat stress responses through ANNEXIN-mediated cytosolic calcium signaling in *Arabidopsis*. *New Phytol.* 216, 163–177. doi: 10.1111/nph.14679
- Ling, Y., Serrano, N., Gao, G., Atia, M., Mokhtar, M., Woo, Y. H., et al. (2018). Thermoprimer triggers splicing memory in *Arabidopsis*. *J. Exp. Bot.* 69, 2659–2675. doi: 10.1093/jxb/ery062
- Liu, Y., Xiao, S., Sun, H., Pei, L., Liu, Y., Peng, L., et al. (2020). AtPRT1, an E3 ubiquitin ligase, enhances the thermotolerance in *Arabidopsis*. *Plants* 9, 1074. doi: 10.3390/plants9091074
- Lobell, D. B., Schlenker, W., and Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science* 333, 616–620. doi: 10.1126/science.12045
- Ma, Y. U., Chen, G. Y., Zhang, L. Q., Liu, Y. X., Liu, D. C., Wang, J. R., et al. (2014). QTL mapping for important agronomic traits in synthetic hexaploid wheat derived from *Aegilops tauschii* ssp. *tauschii*. *J. Int. Agric.* 13, 1835–1844. doi: 10.1016/S2095-3119(13)60655-3
- Mason, R. E., Mondal, S., Beecher, F. W., Pacheco, A., Jampala, B., Ibrahim, A. M., et al. (2010). QTL associated with heat susceptibility index in wheat (*Triticum aestivum* L.) under short-term reproductive stage heat stress. *Euphytica* 174, 423–436. doi: 10.1007/s10681-010-0151-x

- Meng, L., Li, H., Zhang, L., and Wang, J. (2015). QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. *Crop J.* 3, 269–283. doi: 10.1016/j.cj.2015.01.001
- Millet, E. (2008). Exploitation of *Aegilops* species of section Sitopsis for wheat improvement. *Israel J. Plant Sci.* 55, 277–287. doi: 10.1560/IJPS.55.3-4.277
- Mondal, S., Sallam, A., Sehgal, D., Sukumaran, S., Farhad, M., Navaneetha Krishnan, J., et al. (2021). “Advances in breeding for abiotic stress tolerance in wheat,” in *Genomic designing for abiotic stress resistant cereal crops*. Ed. C. Kole (Springer, Cham), 71–103. doi: 10.1007/978-3-030-75875-2_2
- Narayanan, S. (2018). Effects of high temperature stress and traits associated with tolerance in wheat. *Open Access J. Sci.* 2, 177–186. doi: 10.15406/oajs.2018.02.00067
- Neumann, K., Kobylski, B., Denčić, S., Varshney, R. K., and Börner, A. (2011). Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.). *Mol. Breed.* 27, 37–58. doi: 10.1007/s11032-010-9411-7
- 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
- Okamoto, M. (1957). A synaptic effect of chromosome V. *Wheat. Inf. Serv.* 5, 6.
- Ortiz, R., Sayre, K. D., Govaerts, B., Gupta, R., Subbarao, G. V., Ban, T., et al. (2008). Climate change: can wheat beat the heat? *Agric. Ecosyst. Environ.* 126, 46–58. doi: 10.1016/j.agee.2008.01.019
- Palusa, S. G., and Reddy, A. S. (2015). Differential recruitment of splice variants from SR pre-mRNAs to polysomes during development and in response to stresses. *Plant Cell Physiol.* 56, 421–427. doi: 10.1093/pcp/pcv010
- Pankaj, Y. K., Kumar, R., Gill, K. S., and Nagarajan, R. (2024). Unravelling QTLs for non-destructive and yield-related traits under timely, late and very late sown conditions in wheat (*Triticum aestivum* L.). *Plant Mol. Biol. Rep.* 42, 369–382. doi: 10.1007/s11105-022-01357-3
- Patil, I. (2021). Visualizations with statistical details: The ‘ggstatsplot’ approach. *J. Open Source Software* 6, 3167. doi: 10.21105/joss.03167
- Peverelli, M. C., and Rogers, W. (2013). Heat stress effects on crop performance and tools for tolerance breeding. *Rev. la Facultad Cienc. Agrarias* 45, 349–368.
- Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7, e32253. doi: 10.1371/journal.pone.0032253
- Pradhan, G. P., Prasad, P. V. V., Fritz, A. K., Kirkham, M. B., and Gill, B. S. (2012). High-temperature tolerance in *Aegilops* species and its potential transfer to wheat. *Crop Sci.* 52, 292–304. doi: 10.2135/cropsci2011.04.0186
- Qaseem, M. F., Qureshi, R., Shaheen, H., and Shafqat, N. (2019). Genome-wide association analyses for yield and yield-related traits in bread wheat (*Triticum aestivum* L.) under pre-anthesis combined heat and drought stress in field conditions. *PLoS One* 14, e0213407. doi: 10.1371/journal.pone.0213407
- Ramadas, S., Kumar, T. K., and Singh, G. P. (2020). “Wheat production in India: Trends and prospects,” in *Recent Advances in Grain Crops Research*. Eds. F. Shaw, Z. Khan, A. Iqbal, M. Turam and M. Olgun (London: IntechOpen). doi: 10.5772/intechopen.86341
- Ramya, P., Chaubal, A., Kulkarni, K., Gupta, L., Kadoo, N., Dhaliwal, H. S., et al. (2010). QTL mapping of 1000-kernel weight, kernel length and kernel width in bread wheat (*Triticum aestivum* L.). *J. Appl. Genet.* 51, 421–429. doi: 10.1007/BF03208872
- Rasheed, A., and Xia, X. (2019). From markers to genome-based breeding in wheat. *Theor. Appl. Genet.* 132, 767–784. doi: 10.1007/s00122-019-03286-4
- 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/> (accessed January 10, 2022).
- Rife, T. W., and Poland, J. A. (2014). Field book: an open-source application for field data collection on android. *Crop Sci.* 54, 1624–1627. doi: 10.2135/cropsci2013.08.0579
- Riley, R., and Chapman, V. (1958). Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature* 182, 713–715. doi: 10.1038/182713a0
- Rizhsky, L., Liang, H., Shuman, J., Shulaev, V., Davletova, S., and Mittler, R. (2004). When defense pathways collide. The response of Arabidopsis to a combination of drought and heat stress. *Plant Physiol.* 134, 1683–1696. doi: 10.1104/pp.103.033431
- Saini, D. K., Srivastava, P., Pal, N., and Gupta, P. K. (2022). Meta-QTLs, ortho-meta-QTLs and candidate genes for grain yield and associated traits in wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* 15, 1–33. doi: 10.1007/s00122-021-04018-3
- SAS Institute Inc (2013). *SAS/ACCESS® 9.4 Interface to ADABAS: Reference* (Cary, NC: SAS Institute Inc).
- Schneider, A., Molnár, I., and Molnár-Láng, M. (2008). Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. *Euphytica* 163, 1–19. doi: 10.1007/s10681-007-9624-y
- Sears, E. R., and Okamoto, M. (1958). “Intergenic chromosome relationship in hexaploid wheat,” in *Proceedings of the 10th International Congress of Genetics, Montreal*. (Toronto: University of Toronto Press), 258–259.
- Sharma, D., Singh, R., Rane, J., Gupta, V. K., Mamrutha, H. M., and Tiwari, R. (2016). Mapping quantitative trait loci associated with grain filling duration and grain number under terminal heat stress in bread wheat (*Triticum aestivum* L.). *Plant Breed.* 135, 538–545. doi: 10.1111/pbr.12405
- Shiferaw, B., Smale, M., Braun, H. J., Duveiller, E., Reynolds, M., and Muricho, G. (2013). Crops that feed the world 10. Past successes and future challenges to the role played by wheat in global food security. *Food Secur.* 5, 291–317. doi: 10.1007/s12571-013-0263-y
- Shumayla, Sharma, S., Pandey, A. K., Singh, K., and Upadhyay, S. K. (2016). Molecular characterization and global expression analysis of lectin receptor kinases in bread wheat (*Triticum aestivum*). *PLoS One* 11, e0153925. doi: 10.1371/journal.pone.0153925
- Singh, H., Kaur, K., Singh, S., Kaur, P., and Singh, P. (2019). Genome-wide analysis of cyclophilin gene family in wheat and identification of heat stress responsive members. *Plant Gene.* 19, 100197. doi: 10.1016/j.plgene.2019.100197
- Stefanowicz, K., Lannoo, N., and Van Damme, E. J. (2015). Plant F-box proteins—judges between life and death. *Crit. Rev. Plant Sci.* 34, 523–552. doi: 10.1080/07352689.2015.1024566
- Sundeeep, K., Prerna, K., Uttam, K., Monendra, G., Amit, K. S., Rakesh, S., et al. (2013). Molecular approaches for designing heat tolerant wheat. *J. Plant Biochem. Biotechnol.* 22, 359–371. doi: 10.1007/s13562-013-0229-3
- Tahmasebi, S., Heidari, B., Pakniyat, H., and McIntyre, C. L. (2016). Mapping QTLs associated with agronomic and physiological traits under terminal drought and heat stress conditions in wheat (*Triticum aestivum* L.). *Genome* 60, 26–45. doi: 10.1139/gen-2016-0017
- Tao, X., Wang, M. X., Dai, Y., Wang, Y., Fan, Y. F., Mao, P., et al. (2017). Identification and expression profile of CYPome in perennial ryegrass and tall fescue in response to temperature stress. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01519
- Telfer, P., Edwards, J., Norman, A., Bennett, D., Smith, A., Able, J. A., et al. (2021). Genetic analysis of wheat (*Triticum aestivum*) adaptation to heat stress. *Theor. Appl. Genet.* 134, 1387–1407. doi: 10.1007/s00122-021-03778-2
- Tripathi, A., Tripathi, D. K., Chauhan, D. K., Kumar, N., and Singh, G. S. (2016). Paradigms of climate change impacts on some major food sources of the world: a review on current knowledge and future prospects. *Agric. Ecosyst. Environ.* 216, 356–373. doi: 10.1016/j.agee.2015.09.034
- Umezawa, T., Nakashima, K., Miyakawa, T., Kuromori, T., Tanokura, M., Shinozaki, K., et al. (2010). Molecular basis of the core regulatory network in ABA responses: sensing, signaling and transport. *Plant Cell Physiol.* 51, 1821–1839. doi: 10.1093/pcp/pcq156
- Vaid, N., Pandey, P. K., and Tuteja, N. (2012). Genome-wide analysis of lectin receptor-like kinase family from Arabidopsis and rice. *Plant Mol. Biol.* 80, 365–388. doi: 10.1007/s11103-012-9952-8
- Van Berloo, R. (2008). GGT 2.0: versatile software for visualization and analysis of genetic data. *J. Hered.* 99, 232–236. doi: 10.1093/jhered/esm109
- Vinod, J. B., Sharma, J. B., Tomar, R. S., Singh, B., and Tomar, S. M. S. (2012). Breeding for earliness and thermo-tolerance in wheat suited to early, late and very late sowing in northwestern India. *Indian J. Genet.* 72, 15–24.
- Waines, J. G. (1994). High temperature in wild wheats and spring wheats. *Aust. J. Plant Phys.* 21, 705–715. doi: 10.1071/PP9940705
- Wang, J., Liu, W., Wang, H., Li, L., Wu, J., Yang, X., et al. (2011). QTL mapping of yield-related traits in the wheat germplasm 3228. *Euphytica* 177, 277–292. doi: 10.1007/s10681-010-0267-z
- Wang, X., Xin, C., Cai, J., Zhou, Q., Dai, T., Cao, W., et al. (2016). Heat priming induces trans-generational tolerance to high temperature stress in wheat. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00501
- Wei, T., and Simko, V. (2021). R package “corrplot”: Visualization of a correlation matrix (Version 0.92). Available online at: <https://github.com/taiyun/corrplot> (accessed January 26, 2022).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer-Verlag). doi: 10.1007/978-0-387-98141-3
- Wilhelm, E. P., Turner, A. S., and Laurie, D. A. (2009). Photoperiod insensitive Ppd-A1a mutations in tetraploid wheat (*Triticum durum* Desf.). *Theor. Appl. Genet.* 118, 285–294. doi: 10.1007/s00122-008-0898-9
- Zhang, J., Gizaw, S. A., Bossolini, E., Hegarty, J., Howell, T., Carter, A. H., et al. (2018). Identification and validation of QTL for grain yield and plant water status under contrasting water treatments in fall-sown spring wheats. *Theor. Appl. Genet.* 131, 1741–1759. doi: 10.1007/s00122-018-3111-9
- Zhao, W., Liu, Y. W., Zhou, J. M., Zhao, S. P., Zhang, X. H., and Min, D. H. (2016). Genome-wide analysis of the lectin receptor-like kinase family in foxtail millet (*Setaria italica* L.). *Plant Cell. Tiss. Organ Cult.* 127, 335–346. doi: 10.1007/s11240-016-1053-y
- Zhao, J., Missihoun, T. D., and Bartels, D. (2017a). The role of Arabidopsis aldehyde dehydrogenase genes in response to high temperature and stress combinations. *J. Exp. Bot.* 68, 4295–4308. doi: 10.1093/jxb/erx194
- Zhao, Y., Tian, X., Wang, F., Zhang, L., Xin, M., Hu, Z., et al. (2017b). Characterization of wheat MYB genes responsive to high temperatures. *BMC Plant Biol.* 17, 1–14. doi: 10.1186/s12870-017-1158-4
- Zhao, D., Yang, L., Liu, D., Zeng, J., Cao, S., Xia, X., et al. (2021). Fine mapping and validation of a major QTL for grain weight on chromosome 5B in bread wheat. *Theor. Appl. Genet.* 134, 3731–3741. doi: 10.1007/s00122-021-03925-9



OPEN ACCESS

EDITED BY

Ambika Baldev Gaikwad,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Changlong Wen,
Beijing Vegetable Research Center, China
Gung Pyo Lee,
Chung-Ang University, Republic of Korea

*CORRESPONDENCE

Zhuming Tan
✉ tanzm@hnky.cn

[†]These authors have contributed equally to this work

RECEIVED 23 July 2024

ACCEPTED 17 December 2024

PUBLISHED 08 January 2025

CITATION

Shen B, Shen A, Tan Y, Liu L, Li S and Tan Z (2025) Development of KASP markers, SNP fingerprinting and population genetic analysis of *Cymbidium ensifolium* (L.) Sw. germplasm resources in China. *Front. Plant Sci.* 15:1460603. doi: 10.3389/fpls.2024.1460603

COPYRIGHT

© 2025 Shen, Shen, Tan, Liu, Li and Tan. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Development of KASP markers, SNP fingerprinting and population genetic analysis of *Cymbidium ensifolium* (L.) Sw. germplasm resources in China

Baoming Shen[†], Airong Shen[†], Yun Tan, Lina Liu, Sainan Li and Zhuming Tan*

Institute of Biodiversity, Hunan Academy of Forestry, Changsha, China

Cymbidium ensifolium (L.) Sw. is a valuable ornamental plant in the genus *Cymbidium*, family Orchidaceae, with high economic and ecological significance. However, the lack of population genetic information and molecular markers has hindered the development of the sales market and genetic breeding of *C. ensifolium* despite the abundance of commercial cultivars available. In this study, we aimed to develop a set of single nucleotide polymorphism (SNP) markers to distinguish the main cultivated *C. ensifolium* cultivars in China and provide technical support for domestic cultivar protection, registration, and market rights protection. A total of 1,280,516 high-quality loci were identified from 10,021,591 SNPs obtained by sequencing 50 *C. ensifolium* commercial cultivars using double digest restriction site-assisted DNA sequencing technology. A total of 7,599 SNPs were selected for kompetitive allele-specific PCR (KASP) primer design, and 4,360 were successfully designed as KASP markers. Population structure analysis revealed that the 50 commercial cultivars were best divided into four populations, with some correlation between the group distribution and the morphological and geographical characteristics of the germplasm. Using the genotyping results from 28 KASP markers screened from the cultivars, a minimum set of 11 markers was identified that could distinguish 83 *C. ensifolium* commercial cultivars completely, with the remaining 17 markers serving as extended markers. The average PIC value of the 11 markers was 0.345, which was considered medium polymorphism. DNA fingerprints were constructed for the 83 cultivars on the basis of the 11 KASP markers, providing a new approach for mapping DNA fingerprints in *C. ensifolium* cultivars with high efficiency, accuracy, and low cost compared with traditional methods.

KEYWORDS

Cymbidium ensifolium, SNP, KASP, genetic diversity, fingerprint

Introduction

Cymbidium ensifolium (L.) Sw. is a perennial evergreen herbaceous plant in the Orchidaceae family. *C. ensifolium* has a karyotype of $2N = 2X = 40$ with chromosomes of different lengths and the genome size is 3.62 Gb (Ai et al., 2021; Li et al., 2002). It is an important species in traditional Chinese *cymbidium* due to its abundant flower colors, diverse flower types, diverse leaf colorations and unique historical and cultural significance (Cao et al., 2022). *C. ensifolium* has significant ornamental, cultural, medicinal and economic value (Jimoh et al., 2022; Yang and Zhu, 2015). There is a wide variety of *C. ensifolium*, with significant price differences among different varieties (Huang, 2012). However, since *C. ensifolium* varieties are distinguished mainly by flower type and color and some varieties have minimal differences, identification becomes challenging, especially during nonflowering periods (Chen et al., 2022). Traditional morphological identification methods are often insufficient, leading to inaccurate identification and causing discrepancies in orchid trading. This situation adversely affects the robust and orderly development of the orchid market. Additionally, the lack of accurate and reliable identification methods has resulted in confusion between identical *C. ensifolium* varieties with different names and different varieties with the same name, which hampers the identification, conservation, improvement, cultivation, and utilization of *C. ensifolium* genetic resources (Huang, 2012). Therefore, there is an urgent need to establish a simple, stable, and reliable identification method for *C. ensifolium* varieties.

DNA molecular markers have advantages such as short cycle times, minimal environmental influence, and high-throughput detection, providing new means for variety identification (Ramesh et al., 2020). Currently, several molecular marker technologies, including Random Amplified Polymorphic DNA (RAPD), which is based on ITS and cpDNA fragments, Inter-Simple Sequence repeat (ISSR), Simple Sequence Repeat (SSR) markers, and fluorescent SSR markers, have been applied for resource identification of *C. ensifolium* varieties (Li et al., 2014; Jin, 2019; Li, 2014; Wang et al., 2021a; Ai et al., 2019; Hu et al., 2008). However, RAPD is a dominant marker that cannot distinguish between heterozygous and homozygous genotypes, limiting its use in genetic analysis and genetic map construction (Congiu et al., 2000). ITS and limited cpDNA fragment-based markers can only differentiate a small number of varieties because of the limited number of polymorphic sites (Sivu et al., 2022). SSR markers are limited in quantity and detection throughput, have higher detection costs, and require time-consuming and labor-intensive data interpretation, as well as subjective misjudgment of band patterns due to human factors, which restrict their use in a wider range of variety identification work (Zhang et al., 2022b).

Single nucleotide polymorphisms (SNPs) refer to DNA sequence polymorphisms caused by variations such as substitutions and inversions of individual bases in the genomic DNA sequence. SNPs have advantages such as large quantity, wide distribution, allelic dimorphism, and stable inheritance (Li et al., 2023a). These characteristics make SNP-based molecular markers the latest generation of markers. With the emergence of various high-throughput SNP detection platforms, they can effectively

compensate for the technical limitations of SSR markers. The construction of a DNA fingerprint map based on SNP markers is highly important for variety specificity identification, assessment of genetic variations, authenticity verification, identification of seed purity, and other characteristics (Nguyen et al., 2020; Josia et al., 2021; Wang et al., 2021b, 2015; Su et al., 2019). This method has been designated as one of the recommended marker methods by the International Union for the Protection of New Varieties of Plants (UPOV) and the General Guidelines for DNA Identification of Plant Varieties Using DNA markers (NY/T 2594-2016) (Button, 2008). The development of SNP molecular markers can be based on different DNA sources, such as SNP markers developed on the basis of specific genes, EST-SNP markers developed on the basis of express sequence tag (EST) sequences, GSS-SNP markers developed on the basis of gene survey sequence (GSS), and genomic SNPs developed on the basis of whole-genome data (Addison et al., 2020; Wang et al., 2019, 2020; Ashwath et al., 2023). Among these, genomic SNPs provide the most accurate identification but are relatively expensive. With the development of sequencing technology and the popularization of next-generation sequencing (NGS), the cost of sequencing has greatly decreased (van Dijk et al., 2014). Using transcriptome sequencing (RNA-seq), restriction site-associated DNA sequencing (RAD-seq) and double digest restriction site-associated DNA sequencing (ddRAD-seq), more abundant SNP loci can be obtained (Wang et al., 2015; Baird et al., 2008; Magbanua et al., 2023). Among these, RAD-seq and ddRAD-seq can identify SNP loci with broader coverage than can RNA-seq.

There are also various methods for detecting SNP molecular markers, such as direct sequencing, TaqMan probe, amplification refractory mutation system pCR (ARMS-PCR; also known as allele-specific PCR), molecular beacon, high-resolution melting analysis (HRM) technology, cleaved amplified polymorphic sequence (CAPS), SNaPshot, kompetitive allele-specific PCR (KASP), gene chip, and mass spectrometry (Zhao et al., 2017; Kovalchuk and Arkhipova, 2023; Zongze et al., 2018; Gerasimova et al., 2013; Gomes et al., 2018; Amanullah et al., 2022; Pradhan et al., 2023; Guan et al., 2023; Franklin et al., 2020; Wang et al., 2016). Among these, KASP technology has the advantages of low cost, high throughput, time and labor savings, and convenience (He et al., 2014). It has become one of the main methods for SNP genotyping and insertion/deletion (InDel) detection internationally and has been successfully applied in genetic typing and variety breeding of grain crops and economic crops, such as wheat, rice, maize, strawberry, grape, broccoli, cotton, tobacco, and peach (Liu et al., 2023; Tang et al., 2022; Chen et al., 2021; Yang et al., 2020; Wang et al., 2022; Shen et al., 2021; Zhao et al., 2021; Wang et al., 2021b; Fleming et al., 2022). However, there are currently no reports on the application of SNP markers developed on the basis of KASP technology for variety identification, fingerprint map construction, and systematic classification in orchids such as *Cymbidium*.

Therefore, we utilized KASP technology to screen a set of SNP markers that can distinguish *C. ensifolium* germplasm resources in China. In this study, 50 *C. ensifolium* germplasm resources were subjected to Illumina NovaSeq sequencing (Supplementary Table 1). The results were compared with those of the reference genome to identify SNP core markers and interpret their genetic relationships,

genetic diversity, and population structure. Additionally, a DNA fingerprint of 83 *C. ensifolium* varieties was created to effectively differentiate between different varieties of *C. ensifolium* (Supplementary Table 2). These results provide a scientific foundation and data reference for genetic diversity analysis, variety identification, and molecular breeding of *C. ensifolium*.

Materials and methods

Plant materials

Wild *C. ensifolium* resources are included in the National Key Protected Wild Plant List in China, and it is prohibited by law to collect them from the wild. This study adhered to all relevant institutional, national, and international guidelines and laws. No prior permission was required to conduct research on this species. All the plants used in this study are well-known commercial cultivars in China cultivated in pots in a greenhouse under controlled temperature and lighting conditions, a greenhouse on day/night temperatures of 30/23°C under a 14 h light/10 h dark photoperiod. The plant material used in this study was identified by Prof. Zhuming Tan (see author list) and stored at the laboratory of the Hunan Academy of Forestry. To encompass more genetic diversity and obtain a broader range of SNP markers, the selected materials represented diverse morphological variations of *C. ensifolium*, including leaf, sepal, lip, and petal color, size and shape, leaf color and shape, sepal size and shape, lip size and shape, and petal size and shape (Figure 1). Five vigorous plants were randomly selected for each germplasm resource, and young leaf samples were collected, frozen at -80°C, and stored for DNA extraction.

For the preliminary screening of KASP markers, 50 representative *C. ensifolium* cultivars were used (Supplementary Table 1). The rescreening and fingerprint construction using KASP markers included 83 *C. ensifolium* cultivars (Supplementary Table 2).

DNA extraction, library construction and simplified genome sequencing

After the leaf samples were ground, genomic DNA was extracted using a D312 Universal Plant DNA Kit (Genepioneer, China), and the quality and concentration of the DNA were measured, and the purity ($OD_{260\text{ nm}}/OD_{280\text{ nm}}=1.8\text{--}2.0$) was determined using a NanoDrop2000 UV spectrophotometer (Thermo Fisher, Waltham, MA, USA). A paired-end library with a length range of 300–500 bp was subsequently constructed using digest restriction site-associated DNA sequencing library construction of qualified sample DNA (Peterson et al., 2012). First, 500 ng of genomic DNA was incubated with 0.6 U EcoRI (NEB), T4 DNA ligase (NEB), ATP (NEB), and EcoRI connectors (including index sequences of differentiated samples) at 37°C for 3 h and annealed at 65°C for 1 h. The restriction enzyme NlaIII (NEB) and the NlaIII connector were then added and incubated for 3 h at 37°C. After the reaction, the endonuclease was inactivated at 65°C for 30 min in a polymerase chain reaction (PCR) amplifier. The 400–600 bp digested products were recovered using a MiniBEST Agarose Gel DNA Extraction Kit (Takara, China) and quantified using a NanoDrop2000 UV spectrophotometer (Thermo Fisher, Waltham, MA, USA). After 50 samples were mixed in equal quantities, the Illumina NovaSeq 6000 PE150 platform was sequenced and used to construct a DNA library of the mixed products.

Raw data quality control

Raw data quality control was conducted on the original sequenced reads. Fastp software (version: 0.20.0, <https://github.com/OpenGene/fastp>) was used to eliminate reads with an unknown base number $N<5$, a quality value <5 , connector sequences, and other low-quality sequences to obtain clean data. Burrows–Wheeler Aligner (BWA) (version: 0.7.17-R1188, <https://github.com/lh3/bwa>) was then used to align the sequenced reads with the reference genome (Li and

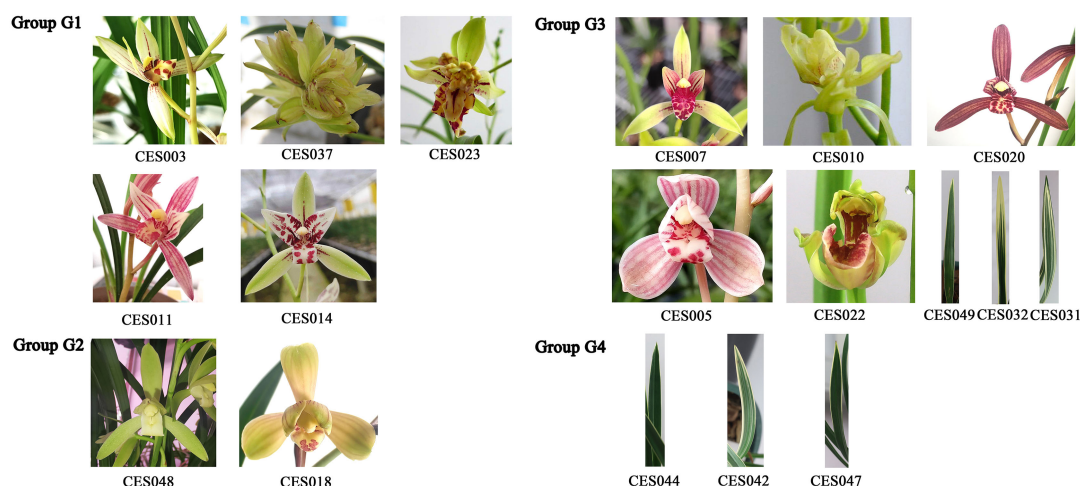


FIGURE 1
Representative flower and leaf characteristics of the tested *C. ensifolium* varieties.

Durbin, 2009). The reference genome used was the GWHBCII000000000 genome Fasta (*C. ensifolium*) (download website: <https://ngdc.cncb.ac.cn/gwh/Assembly/20686/show>), with parameters set as -M -R [51]. The insert size and coverage depth of each sample were determined, and variation was detected by comparing the positions of the clean reads in the reference genome (Li et al., 2023a). SAMtools software (version: v1.9, <https://github.com/samtools/samtools>) was used to statistically analyze the sequencing depth, genome coverage, insertion fragment length, and other information of each simplified genome sample (Li et al., 2009).

SNP analysis

SNP analysis was primarily performed using the Genome Analysis Toolkit (GATK) software package (version: 4.1.4.1, <https://github.com/broadinstitute/gatk>). Using the positioning of clean reads in the reference genome, GATK was used to detect SNPs and obtain the final SNP site set, followed by SNP statistics (McKenna et al., 2010). The main detection process included the following steps: (1) Picard's (version: 0.7.17-r1188, <https://github.com/broadinstitute/gatk>) Mark Duplicate tool was used to remove duplicates and mask the effects of PCR duplication on the results from BWA alignment; (2) variant calling was performed using GATK, including SNP and InDel; and (3) variant quality score recalibration (VQSR) was performed using GATK. (4) GATK was used to filter the obtained variants (filtering out sites with QD < 2.0, MQ < 40.0, FS > 60.0, SOR > 6.0, MQRankSum < -12.5, ReadPosRankSum < -8.0), selecting reliable variant results, followed by statistical analysis and bar chart plotting of the variant types by GATK. SNP site annotation was implemented with SnpEff (version: 4.3t, <http://pcingola.github.io/SnpEff>).

Population SNP filtering

To obtain high-quality SNPs for population genetic analysis of *C. ensifolium*, a series of standards were applied for preliminary screening. These standards included criteria such as average sequencing depth $\geq 5\times$, minor allele frequency (MAF) ≥ 0.05 , information integrity ≥ 0.70 , SNP quality value $Q \geq 30$, and number of alleles of 2 (Li et al., 2023a). The SNP density distribution across each chromosome was subsequently calculated and visualized. The average values of the population genetic indices in 50 *C. ensifolium* samples were also calculated and statistically analyzed.

Genetic diversity and population structure analysis

Genetic diversity and population structure analyses were performed using GCTA software (version: 1.92.1, <http://cnsngomics.com/software/gcta/#Overview>) for principal component analysis (PCA) on the basis of the high-quality SNPs obtained. The maximum likelihood (ML) method in RAxML

software (version 8.2.12, <https://github.com/stamatak/standard-RAxML/>) was used to construct an evolutionary tree of 50 *C. ensifolium* samples. Admixture software (version: 1.3.0, <http://software.genetics.ucla.edu/admixture/>) was used to analyze the population genetic structure.

SNP site filtering and KASP primer design

Due to the difficulty in distinguishing between samples in this study, a combination of conventional filtering and manual selection of difficult-to-distinguish varieties was used in the SNP site selection strategy. After considering both filtering criteria, the selection criteria for SNP sites used in KASP primer development in this study were defined as follows: first, conserve the sequence flanking the SNP site on the DNA chain of the chromosome for more than 50 bp; second, retain markers with an average depth of 5X or greater, with the SNP being a biallelic gene; third, trim 100 bp sequences upstream and downstream of the SNP marker, then use blast software (version: 2.10.1+, <https://blast.ncbi.nlm.nih.gov/Blast.cgi>) to align the sequence against the reference genome and remove markers with multiple alignment positions; fourth, retain markers with a polymorphic information content (PIC) greater than 0.020, (the low PIC threshold is due to manual selection considerations), and the PIC calculation formula refers to the method of Zhang, et al., where P_i and P_j are the frequency of occurrence of the two alleles of SNP in all varieties tested, and l is the number of samples (Zhang et al., 2020).

$$PIC = 1 - \sum_{i=0}^l P_i^2 - \sum_{i=1}^{l-1} \sum_{j=i+1}^l 2P_i^2 P_j^2$$

For each selected SNP marker, primer design was conducted using Primer3 software (version: 2.4.0, <https://sourceforge.net/projects/primer3/files/latest/download>). Sequences 100 bp upstream and downstream of candidate SNPs were used for KASP marker design. Two allele-specific primers and one universal primer were designed for each KASP target site. The primer design parameters were set as follows: GC content < 60%, melting temperature (T_m) between 57 and 63°C, and PCR product size not exceeding 120 bp. Further optimization of primer design results can be conducted on the basis of actual verification results. Only SNP sites with successful primer design were considered to be qualified KASP markers and were used for downstream analysis. The distribution statistics of the markers with successful primer design were also based on the annotation results.

Verification of SNP locus authenticity and KASP genotyping

To save costs, 28 randomly selected SNP sites with successful KASP primer design were preliminarily screened using Perl scripts, with 5 samples selected for first-generation sequencing validation. The distribution of developed KASP markers on the chromosomal was visualized using Tbttools-II (version: 2.138) (Chen et al., 2023).

Primers were designed, conventional sequencing was performed, and the authenticity of the SNP site was verified

TABLE 1 Primers used for the initial screening validation of candidate SNP loci.

Number	SNP Position	Forward primer (5'-3')	Reverse primer (5'-3')	Product length/bp
1	Chr1-55623523	AGCGTGGTGGGATTGAACCT	ACGAAAGCACGACGTGTACA	242
2	Chr1-73862970	GATTGTCCTATTCAAGTCTCATTAGA	ACCTTTCTTGGCTTGGCAAA	164
3	Chr1-171445320	ACCAAACGTGTAGGCAACCA	CCCTCGACCGTAAGTTCAC	196
4	Chr10-25134624	CGGAGCAAACCTCATCACCT	GCAACACACCAGAGGTACCA	191
5	Chr12-53446226	CTCTCGAGGAGCTATTGCAG	GCCCACTAAAGTGCAAAGCA	495
6	Chr12-63867683	AGAAGCGATCAAGAACCGGG	AGCATGTCTCACACTGGT	278
7	Chr12-75662664	TTAGATGGGCCAGGCTTGTG	GGTGCATGGTGAGTTCAGCT	547
8	Chr13-6407723	GGATGGCTCTGCTTCAGTT	CCTAACTACGCCCTTCGCAA	364
9	Chr13-34339702	TCAGTTTCGGACGAAGGTCG	ATGTTTGTGCCACGCGTTTC	271
10	Chr16-43018260	CTATGGGCGTAGAAGTGGTGT	AGTGTCAACATTGCGGCAAC	367
11	Chr16-66004509	GGCGCATCCAAAGAACAAGG	CCTCTAATATGATCTTGAAGTGCACC	308
12	Chr18-11319799	CCAAGAACCCTCATGTCCC	TGAGCCATGATAATTGCAAGGT	381
13	Chr18-27339793	TCGCTTATCAGAAGTGCCACA	TGCTGTTTACTGCCCATCCA	532
14	Chr19-8308837	TGGCAGAAACCTATAGGCCT	TTCCCTGCCAAGCTCAATGT	177
15	Chr2-12538713	ACGGTTGCTCGGCTTTTCTA	ACCATCCAAACCGTAAGCTTCT	329
16	Chr2-77710898	GTCCCAAAGACAAGTTCCCA	GGCTTGCTAATACGTGGCAC	221
17	Chr2-114570305	TGCAAGCAAGATGTCTCCGT	CCACGCACTTACAGGAGGTT	244
18	Chr2-140839794	CGGAGCTTCAATAGTTAACAACAGT	TCTTTCTGATGCAGGGCTGG	475
19	Chr20-7711682	TGGTTTCCACCTTGCCTGTT	CTGATGGCTCCTGGTGTGTT	208
20	Chr20-9861983	TGAATATACATTTCAACCATGGAGTT	TTTTCCTTGCAGAGACTTGC	237
21	Chr3-67201341	TCGCTCCTTCGCTTGTTTCAT	CAAGCAGCCCTTGGTATTGA	549
22	Chr3-153114895	GCCAAGCCTGAGGTGTTAGA	TGTGAGTCAGCTTGCACAGT	190
23	Chr4-57124129	CTCCACCACCTGCATCCATT	TGGATATCCATATGAAGGTGTTGCA	440
24	Chr4-103337174	TCCTGGTTCCCTGATCCCAT	GATAATGGGGCCGAGAAGG	471
25	Chr5-47997902	TGCATCCTAACAAGACTTGGCA	ACAGTCAGTTAGAGCGCAACA	500
26	Chr8-90715849	GAGTCAGGTTTCCAAACGCG	AACATGTCGCGGTACAACCT	335
27	Chr9-22416891	GGACTTCAAGGAAGTGGAGGA	AGAAAGAAACCGCGCAAAC	511
28	Chr9-24025081	ACCTTGTTGGCCGCCTTTTA	AAGCCGAGCTGCAATAAACCC	431

Each SNP location information consists of the number of chromosomes where the SNP site is located plus the location of the corresponding chromosome.

(Table 1). For successfully validated sites, KASP primers were redesigned using the real first-generation sequencing results with Primer3 (version: 2.4.0, <https://sourceforge.net/projects/primer3/files/latest/download>). The primers were synthesized by Sangon Biotech (Shanghai) Co., Ltd., with FAM- or VIC-tails (FAM-tail: 5'-GAAGGTGACCAAGTTCATGCT-3'; VIC-tail: 5'-GAAGGTCGGAGTCAACGGATT-3') (Supplementary Excel 1). Subsequently, KASP genotyping experiments were subsequently conducted.

The KASP assay was performed in a 5 μ L PCR system/condition comprising 2.5 μ L of 2X KASP Master mix (JasonGen, China), 1.25 μ L of primer mixture, and 1.25 μ L of DNA at a concentration of 10–20 ng/mL. The PCR program included 10 min at 95°C, 10 touchdown cycles of 95°C for 20 s and 61–55°C for 60 s (decreasing by 0.6°C per cycle),

and 27 cycles of 95°C for 20 s and 55°C for 60 s. Following PCR, fluorescence data were read and analyzed via the CFX Connect™ Real-Time System (Bio-Rad, USA). If the genotyping results were unsatisfactory, PCR optimization was performed with an additional 3 cycles of 95°C for 20 s and 55°C for 60 s.

Fingerprint construction

For the obtained SNP genotyping results, the optimal combination of markers was calculated for fingerprinting using a Perl program. The genotypes of the optimal combination of markers were heatmapped for fingerprinting, with each row representing one SNP locus and each column representing one

sample. The genotypes were color-coded as follows: AA=green, AG=light pink, CC=yellow, CT=grew, GT=dark red, TT=blue, AC=pink, AT=orange, CG=light blue, and GG=purple; and no call genotypes were designated as NN=white.

Results

Simplified genome sequencing and reference genome alignment

The sequencing of 50 *C. ensifolium* samples on the Illumina NovaSeq 6000 PE150 yielded a total of 64.89 Gb of clean data, with an average sequencing data volume of 1.30 GB per sample. The Q30 scores ranged from more than 89.79% to 94.26%, with an average exceeding 92.76%. The average Q score was between 35.24 and 36.07, and the GC content ranged from 33.86% to 35.17% (Supplementary Table 3). The clean reads obtained were mapped to the reference genome, with an average mapping efficiency of 99.50%. The genome average cover depth was 6.6X, the genome coverage at 5X averaged 1.92%, and that at 10X averaged 0.95% (Supplementary Table 4).

Selection, identification and annotation of high-quality SNPs

Following sequencing, GATK software detected numerous SNP variants in the 50 *C. ensifolium* samples, resulting in a total of 10,021,591 SNPs, with each sample ranging from 691,307 to 1,612,362 (Supplementary Table 5). The most common types of SNP variations were C>T (2,034,818) and G>A (2,037,584), followed by A>G (1,289,885) and T>C (1,286,472), with the least common types being C>G (233,292) and G>C (233,361) (Supplementary Figure 1A, Supplementary Excel 3). A total of 1,280,516 filtered SNPs distributed across 20 main chromosomes were obtained for population genetic analysis (Supplementary Figure 1B, Table 2). An SNP distribution map was then created for the 20 main chromosomes on the basis of the number and density of SNPs (Figure 2). Chromosome GWHBCII00000001 had the highest number of SNPs (67,609), whereas GWHBCII000000020 had the lowest number of SNPs (14,535). The SNPs were evenly distributed on the 20 main chromosomes, with an average of 322 SNPs per MB (data not shown).

Genetic relationships and population structure analysis

Admixture software was used to analyze the population structure of the 50 *C. ensifolium* samples. The population was

divided into four subgroups (G1, G2, G3, G4) on the basis of the lowest cross-validation error rate at K=4 (Figure 3A). G2 included 5 varieties, which primarily descended from a common ancestor and predominantly represented purebred varieties. G4 comprised 8 varieties, predominantly hybrids, including several leaf art varieties. The G1 and G3 subgroups were more diverse, encompassing a variety of flower colors, patterns, and leaf art varieties. G1 comprised 14 varieties, with only CES011 descending from 3 ancestors, while the rest originated from a common ancestor. G3 included 23 varieties, with 16 having a single ancestral origin (Figures 3B, D).

PCA was conducted using GCTA software on high-quality SNPs from the 50 *C. ensifolium* samples, resulting in the samples being discriminated into four groups on the basis of the first two components, accounting for 36.99% and 6.79% of the total variation, respectively (Figure 3C).

An evolutionary tree was also constructed using RAxML software, which clustered the 50 *C. ensifolium* samples into four groups (Figure 4), which was consistent with the PCA and population structure analysis results and was supported by high bootstrap values. The clusters presented a certain degree of similarity in terms of phenotype and origin (Supplementary Table 1).

Development and verification of highly polymorphic KASP SNP markers

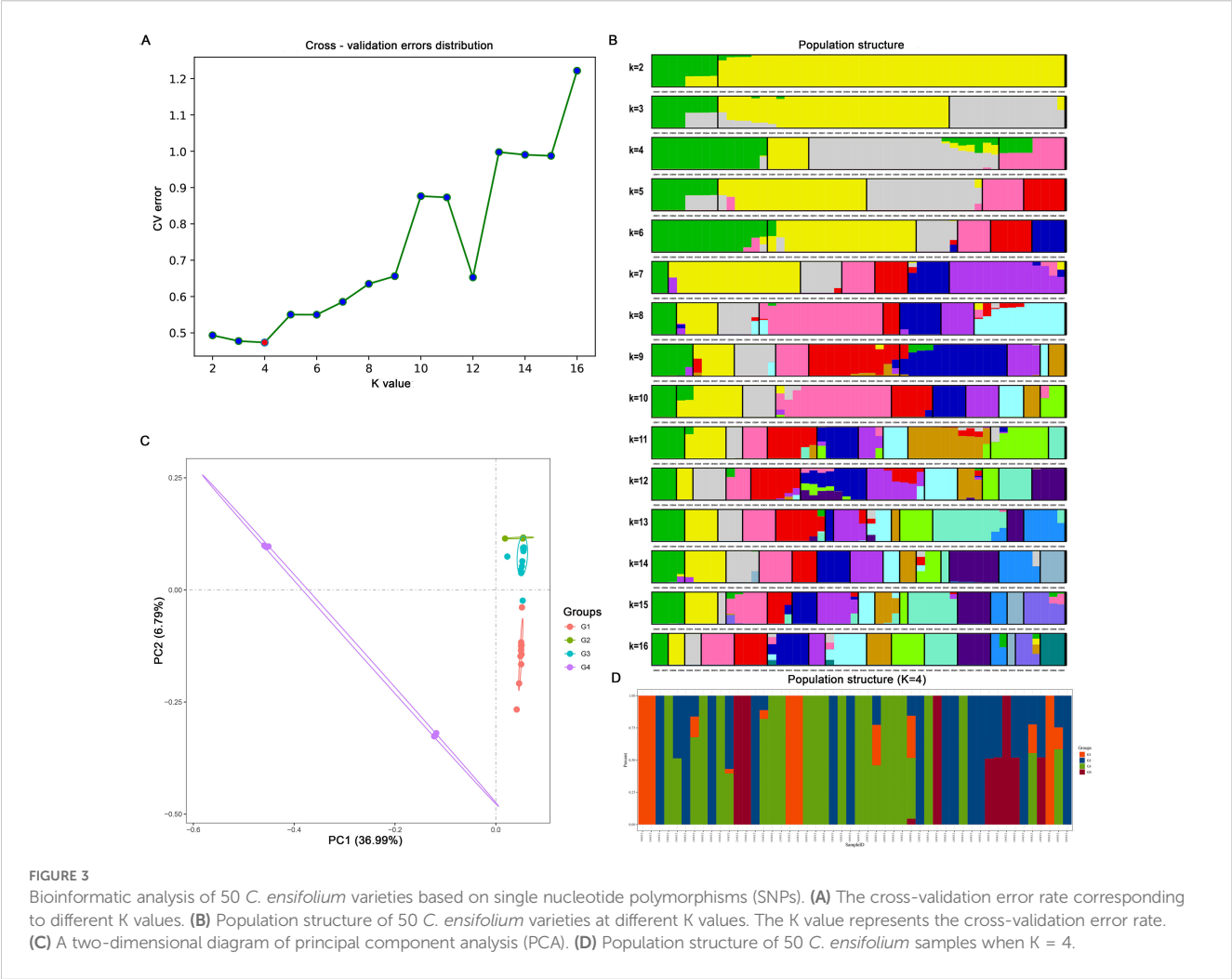
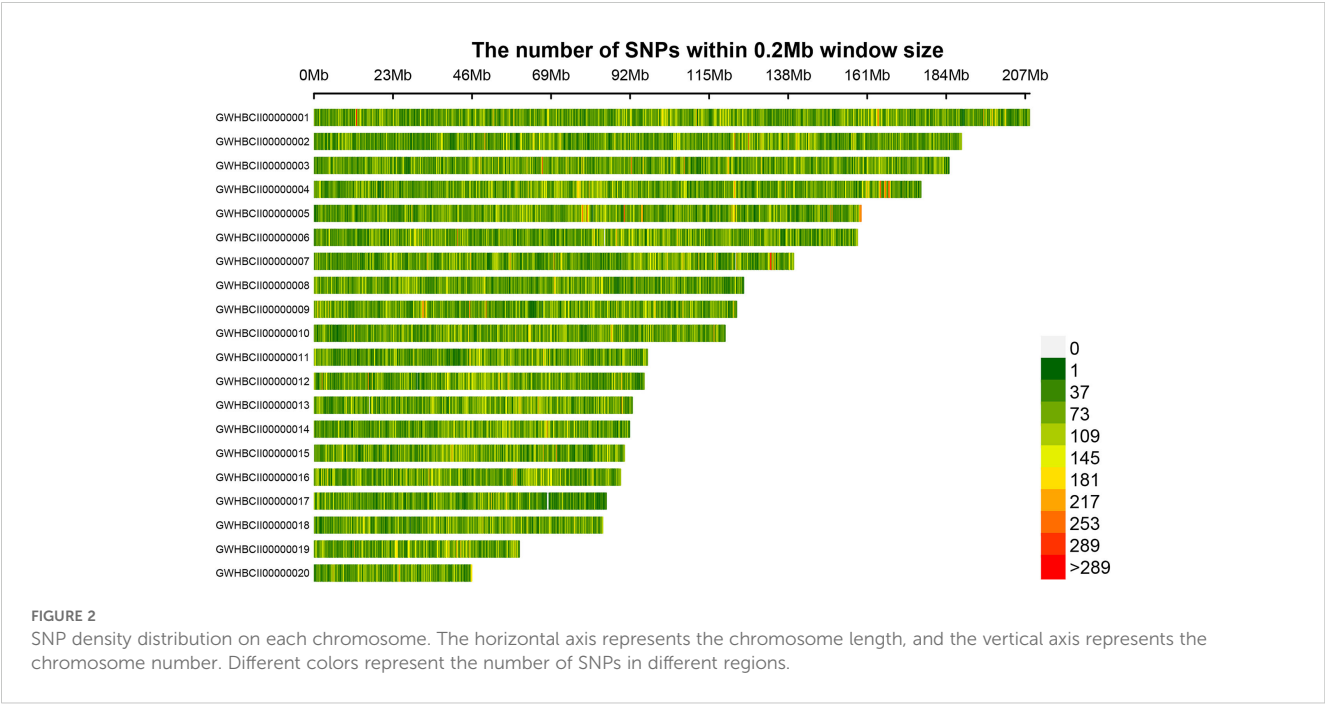
To develop high-quality and polymorphic KASP SNP markers that could effectively differentiate the tested samples, stringent criteria were applied, including SNP conservation and uniqueness on the chromosome DNA strands, high sequencing depth, biallelic polymorphisms, and high PIC values. A total of 7,599 SNPs were screened from 10,021,591 SNPs, with 4,360 successfully designed as KASP markers (Supplementary Excel 2). Among these KASP markers, 428 were located in exonic regions, accounting for 9.8% (Supplementary Figure 2).

To validate the practicality of the selected SNP markers, we utilized a Perl program to screen 28 loci from the 4,360 SNPs mentioned above for verification (Supplementary Excel 4, Supplementary Figure 3). The first-generation sequencing results indicated that all 28 loci were indeed present (Table 1). On the basis of the actual sequencing results, we subsequently optimized the primer design for KASP and conducted KASP genotyping experiments (Supplementary Table 3). Among the 50 simplified sequencing samples, most samples were well genotyped and effectively distinguished, but two groups (CES001, CES002, CES019, CES048; CES035, CES043) could not be differentiated (data not shown). Despite multiple rounds of screening, these two

TABLE 2 Average population genetic index values for the 50 *C. ensifolium* varieties genome simplification sequencing samples within the population.

Population	Sites	Polymorphic	HetObs	HetExp	Pi	Fis	PIC	Ne	MAF
All	1,280,516	1,280,516	0.203	0.270	0.270	0.249	0.225	1.423	0.186

The population genetic index values are averages. Population, population ID; Sites, number of SNP markers within the population; Polymorphic, number of polymorphic markers; HetObs, observed heterozygosity; HetExp, expected heterozygosity; Pi, nucleotide polymorphism within the population; Fis, average inbreeding coefficient of the population; PIC, polymorphic information content; Ne, effective number of alleles; MAF, minimum allele frequency.



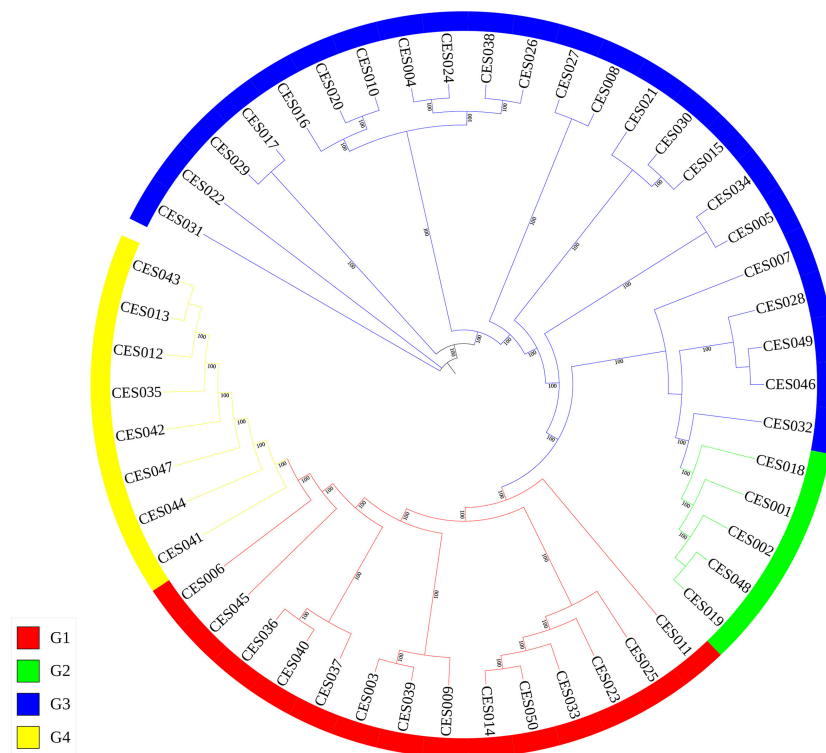


FIGURE 4
Phylogenetic tree of 50 *C. ensifolium* varieties. Branches of the same color are in the same group.

groups could not be distinguished, making the development of KASP markers difficult. Considering the cost of developing markers, this study did not continue to develop KASP primers to differentiate these similar varieties. To verify the utility of the selected KASP markers, we selected another 37 cultivated varieties of *C. ensifolium* for KASP genotyping. The results revealed that four markers (Chr12-75662664, Chr16-66004509, Chr18-11319799, Chr3-153114895) had only one genotype, whereas the other 24 markers were polymorphic and could effectively distinguish the other 37 cultivated varieties of *C. ensifolium* (Figures 5A, B; Supplementary Excel 5). These findings indicate that the selected KASP markers have good usability. Due to the large number of *C. ensifolium* varieties, by expanding the screening range these four KASP primers with consistent genotypes may still exhibit polymorphisms, thus supporting the potential value of the markers developed in this study.

Construction of a DNA fingerprint

To protect the genetic resources of *C. ensifolium* varieties, we constructed a fingerprint map for all tested varieties on the basis of the KASP marker genotyping results, with one variety selected from each undistinguished group as a representative (CES001, CES035). A good fingerprint map requires a minimum number of markers to distinguish the maximum number of varieties to achieve simplicity, efficiency, and economy. By selecting 11 SNPs from the 28 obtained SNP loci that could distinguish 83 *C. ensifolium* varieties, we

constructed a highly simplified SNP combination distributed across seven chromosomes (Figure 5C). Using these 11 SNPs, we successfully distinguished all 83 varieties, and there was at least one different SNP between each of the two samples. The PIC values of the 11 core KASP SNPs ranged from 0.19 to 0.37, and 90% > 0.33, indicating medium genetic diversity (Supplementary Excel 4). The combination of these 11 SNPs effectively distinguished the *C. ensifolium* varieties, and the fingerprint map constructed on the basis of the genotyping results further confirmed this distinction (Figure 5D). The remaining 17 KASP SNP markers can be used as backup markers when expanding sample testing to complement and improve the fingerprint map (Supplementary Excel 4).

Discussion

SNP-based genetic relationships among the *C. ensifolium*

C. ensifolium is widely distributed in southern China and is highly diverse in terms of flower color, flower type, leaf color, leaf shape, and plant morphology. As a traditional Chinese *Cymbidium* species, *C. ensifolium* is highly valued for its historical, economic, and ornamental significance (Cao et al., 2022; Ai et al., 2021). In the past, wild types with unique traits were sought from the field to breed better varieties of *C. ensifolium*. However, due to the emphasis on the protection of orchid species in China, the collection of wild genetic resources has been prohibited by law. Therefore, breeding new

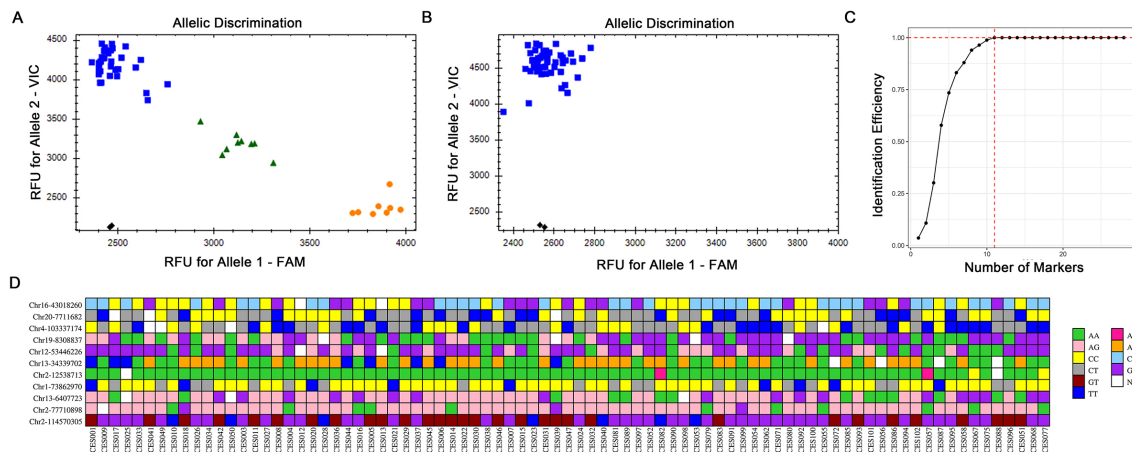


FIGURE 5

Representative KASP-labelled fluorescence assay results and fingerprint analysis of 83 *C. ensifolium* varieties. (A) A KASP marker with better typing. (B) A KASP marker with monomorphism. (C) Identification efficiency of the combined SNP markers. (D) Fingerprints of the 83 *C. ensifolium* materials. Each row represents the typing result of the same label in different samples, and each column represents a sample. The pure genotypes were AA=green, AG=light pink, CC=yellow, CT=grew, GT=dark red, TT=blue, AC=pink, AT=orange, CG=light blue, and GG=purple; and no call genotypes were designated NN=white.

varieties from existing varieties has become the primary means of creating new orchid varieties. Current studies have focused mainly on the genetic control of specific traits in plants, such as flower color, leaf color, and drought resistance, providing references for future breeding techniques (Ai et al., 2023; Li et al., 2023b; Mei et al., 2024). Before new varieties are bred, it is essential to determine the genetic and kinship relationships between breeding materials (Chen et al., 2020). However, few studies have investigated the genetic and kinship relationships among commercial *C. ensifolium* varieties at the genomic level.

In this study, 50 mainstream commercial varieties of *C. ensifolium* were selected, and using ddRAD-seq technology, a series of polymorphic SNP markers were obtained through strict screening criteria. Genetic and kinship analyses of the 50 commercial varieties of *C. ensifolium* on the basis of SNP sequences were conducted. PCA and population structure analysis indicated that the 50 *C. ensifolium* varieties could be divided into 4 groups, which was also supported by the results of the maximum likelihood phylogenetic tree. The selected *C. ensifolium* varieties in this study overlap with those selected by Ai et al., but the clustering results are different, which may be due to different population constructions and segments selected for building the phylogenetic tree (Ai et al., 2019). Among the 50 selected *C. ensifolium* varieties, most originated from Taiwan, such as *C. ensifolium* var. 'Qi Xiannv', *C. ensifolium* var. 'Shizhang Hong', *C. ensifolium* var. 'Lishan Shiwang', *C. ensifolium* var. 'Hong Niang', *C. ensifolium* var. 'Jin He', *C. ensifolium* var. 'Fushan Qidie', and *C. ensifolium* var. 'Baodao Xiannv', which were all classified in Group 1 (G1), whereas those originating from Sichuan were mostly classified in Group 3 (G3). This finding indicates that the grouping in this study is consistent with the original sources of the varieties. Even though varieties from different regions have undergone long-term domestication, they still retain their own distinct features (Lin et al., 2022). Within the G2 group, most individuals are from

the southeastern coastal regions of China. In addition to the hybrid variety *C. ensifolium* var. 'Zhu Jin', the other varieties also share some common phenotypic characteristics, such as flowers with light-colored sepals and labellums that are solid in color without spots. On the other hand, the G4 group consists mainly of hybrid species. Additionally, some cultivars were not grouped together with others from the same geographical origin, indicating widespread genetic exchange among cultivated varieties.

Hybrid breeding involves selecting parental varieties with a large genetic distance and distant kinship, which aids in selecting new varieties with excellent traits (Geng et al., 2021; Cui et al., 2023). The evolutionary tree in this study shows clear genetic distances and kinship relationships among the 50 *C. ensifolium* varieties. For example, *C. ensifolium* var. 'Zhaojun Xue', *C. ensifolium* var. 'Da Jiangjun', *C. ensifolium* var. 'Huangjin Xianzi', and *C. ensifolium* var. 'Danxia Xiannv' are located on the same branch and have large genetic distances from the remaining varieties, providing theoretical references for the selection of parental varieties in *C. ensifolium* breeding.

Cultivar identification and the advantages of SNP molecular markers

C. ensifolium is a highly valuable ornamental and economic orchid, and through generations of breeding, there are currently an estimated thousand varieties of *C. ensifolium* available on the market (<https://www.hmlan.com/auction/search-101005.htm?q=%BD%A8%C0%BC&noex=>) (accessed on 5 June 2024). Identifying the correct variety of orchids to purchase online during nonflowering periods has become a challenge for many orchid enthusiasts. The traditional classification of *C. ensifolium* varieties is based mainly on morphological characteristics. Wang used 14 quantitative traits and 11 qualitative traits to describe 39

Cymbidium varieties in detail, providing a phenotypic basis for the identification of these varieties (Wang, 2020). Nevertheless, identifying varieties during nonflowering periods remains difficult. To address the shortcomings of morphological identification, researchers have also attempted to use various molecular methods for *C. ensifolium* variety identification. For example, Wang et al. successfully differentiated 9 orchid species, including *C. ensifolium*, using ALFP technology (Wang and Wang, 2014). Wang et al. also distinguished 85 *C. ensifolium* cultivars using 19 ISSR primers (Wang et al., 2011). Hu et al. differentiated 38 *C. ensifolium* varieties using 18 RAPD primers with genetic distances ranging from 0.0420 to 0.5385 [12]. Li et al. developed 55 genic-SSR polymorphic markers from the total RNA of *C. ensifolium* var. Tiegusu and distinguished 9 *Cymbidium* species and 12 *C. ensifolium* cultivars using evolutionary tree construction. The genetic distance ranged from 0.016 to 0.618 (Li et al., 2014). Although these methods can be used to differentiate *C. ensifolium* cultivars, they have limitations such as inconvenient operation, high cost, inability to distinguish genotypes, and low number of polymorphic sites, highlighting the urgent need for a new, simple, accurate, and efficient molecular identification method.

SNP molecular marker technology has numerous advantages, including a large quantity, wide distribution, allelic dimorphism, and stable inheritance (Li et al., 2023a). This technology has been successfully applied in various crops, such as rice, grape, potato, cotton, radish, and honeysuckle (Xing et al., 2024; Wang et al., 2022; Li et al., 2023a; Morales et al., 2020; Gazendam et al., 2022; Kuang et al., 2016). It is one of the marker methods recommended by the UPOV and the general guidelines for the identification of plant varieties using DNA markers (NY/T 2594-2016) (Button, 2008). However, its application in *Cymbidium* species, particularly *C. ensifolium*, is relatively limited (Yang et al., 2023).

In this study, using the ddRAD-seq technology on 50 *C. ensifolium* cultivars, a total of 10,021,591 SNP loci were obtained, surpassing the number of polymorphic markers obtained by Li et al. using genic-SSR markers in *C. ensifolium* (Li et al., 2014). On average, 964,159 SNP loci were developed per cultivar, whereas 334,967 SNP loci were obtained per *Cymbidium sinense* (Jack. ex Andr.) Willd. cultivar using SLAF-seq technology (Yang et al., 2023). These results indicate the presence of a rich SNP locus population in orchids, providing a valuable resource for developing SNP markers. After stringent filtering, 4,360 highly polymorphic SNP loci were selected, and a set of 11 SNP markers were identified that could effectively distinguish 83 *C. ensifolium* cultivars. This identification efficiency is significantly greater than that of traditional *C. ensifolium* classification methods such as AFLP, ISSR, RAPD, and genic-SSR (Wang and Wang, 2014; Wang et al., 2021a, 2011; Hu et al., 2008; Li et al., 2014).

The main advantages of the KASP genotyping technique for identifying varieties

The Laboratory of Government Chemists (LGC) developed a high-throughput genotyping technique based mainly on SNPs

based on the principle of KASP (Dipta et al., 2024). Overall, the KASP genotyping technique offers several advantages for identifying varieties. Compared with other SNP genotyping technologies, it is cost-effective, with lower material expenses per reaction (Ayalew et al., 2019; Yuan et al., 2014). Additionally, it is a simple and gel-free assay that can be easily performed using regular qPCR instruments, reducing labor costs. Compared with AFLP and SSR technology, the automatic genotyping technique, which is based on fluorescence differences, also minimizes the effect of error when the gel electrophoresis image is read (Wang and Wang, 2014; Sun et al., 2023). The design principle behind KASP primers allows quick and accurate genotyping on the basis of SNP polymorphisms, making it a valuable tool for quality control and QTL mapping (Kumar et al., 2022; Zeng et al., 2022). In conclusion, the KASP genotyping technique is a reliable and efficient method for identifying genetic variations in plant varieties.

Due to the numerous advantages of the KASP genotyping technique, KASP assays have been developed for genotyping analysis in a variety of plant species. For example, in conventional and hybrid rice, Tang et al. developed 48 KASP markers, and the 48 KASP markers had a 100% discrimination rate in 53 conventional indica varieties and 193 hybrid varieties (Tang et al., 2022). In cabbage, Li et al. selected 442 KASP SNP markers among 50 resequenced genotypes on the basis of high polymorphism information content, high minor allele frequency, wide average distribution and low heterozygosity. Using the KASP genotyping data, the genetic similarity among three kinds of inbred lines (spring cabbage, autumn cabbage and winter cabbage) was analyzed, and the heterotic groups within each ecotype were classified. Seven heterotic groups were identified for spring cabbage (77), six for autumn cabbage (70), and five for winter cabbage (97) (Li et al., 2020). In *Brassica rapa*, Hong et al. developed 100 accession-specific markers as accession-specific KASP markers. Using the results of their validation experiments, they successfully distinguished the accession-specific markers in individual accessions in test populations from noncore or commercial cultivars (Hong et al., 2022). In addition, KASP markers play important roles in genotyping and variety identification in other plants, such as apple, coffee, and wheat (Winfield et al., 2020; Patterson et al., 2017; Roncallo et al., 2019; Akpertey et al., 2020; Zhang et al., 2023b; Kumar et al., 2022). However, little research has been conducted on SNP markers and KASP marker development in *C. ensifolium*.

In our study, using ddRAD genome simplified sequencing data of 50 *C. ensifolium* commercial varieties, we obtained 10,021,591 SNP sites. After a series of strict screening criteria, we obtained 7,599 SNP sites with high polymorphism information content, high minor allele frequency, and a wide average distribution. Among these, 4,360 SNP markers were successfully converted into KASP markers, accounting for 57.4% of all 7,599 SNP markers. The conversion efficiency of KASP markers was lower than that of Cabbage (88.4%), *Capsicum annuum* L. (88.2%), and rice (94.8%) (Zhang et al., 2023b; Li et al., 2020; Yang et al., 2019). In terms of the conversion efficiency of all SNP sites into KASP sites, this study revealed an efficiency of 0.04%, which was lower than that of maize (2.42%) (Chen et al., 2021). These findings indicate that the development of KASP SNP markers for *C. ensifolium* is relatively

challenging. These KASP SNP markers provide a rich resource database of polymorphic sites for the identification and consistency testing of *C. ensifolium* varieties and even orchid varieties.

Identification and validation of 83 *C. ensifolium* germplasm resources by SNP fingerprints

Germplasm resources, as important biological resources, constitute the genetic basis for breeding high-quality and unique new varieties (Wang et al., 2021b; Li et al., 2023a). With the development of technology, the types and quantity of discovered germplasm resources are becoming increasingly diverse and abundant. Conducting DNA fingerprinting on germplasm resources is a good choice for effective protection and efficient utilization of many germplasm resources. DNA fingerprinting was first proposed by the geneticist Alec Jeffreys from the University of Leicester in 1985. DNA fingerprinting uses isolated human microsatellite DNA as a gene probe, hybridizes it with enzyme-cut fragments of human nuclear DNA, and obtains hybrid bands composed of alleles from multiple loci of different lengths. These fingerprints are unique to each individual, similar to human fingerprints (Jeffreys, 2013). The construction of a DNA fingerprint map can provide each germplasm with a unique identity, assist in accurate identification of germplasm resources, and play an important role in variety specificity and authenticity, seed purity identification, improved resource utilization efficiency, and protection of the intellectual property of plant breeding (Xing et al., 2024; Shen et al., 2021; Yang et al., 2022; Tian et al., 2021). SNP molecular markers are widely used for constructing DNA fingerprint maps in various plants, such as maize, cucumber, honeysuckle, cigar tobacco, radish, and red raspberry (Mannino et al., 2023; Zhang et al., 2024; Wei et al., 2024; Li et al., 2023a; Wang et al., 2021b; Tian et al., 2021; Xing et al., 2024; Clare et al., 2023; Zhang et al., 2022a). The accuracy of SNP fingerprinting has been validated in cucumber through distinctness, uniformity, and stability (DUS) testing, further demonstrating the accuracy and practicality of SNP fingerprinting (Zhang et al., 2022a). However, there have been no reports of SNP fingerprints of *C. ensifolium* germplasm resources.

SNP molecular markers are widely used for gene identification, germplasm characterization, and variety fingerprinting (Yang and Zhu, 2015; Chen et al., 2022; Magbanua et al., 2023; Zhao et al., 2017). However, it has been reported that only a small proportion of SNP loci can be selected and genotyped successfully. The authenticity of SNP sites in simplified floral genome sequencing is low, making the validation of SNP site authenticity even more important for constructing a fingerprint map (Liu et al., 2022a). Our study also revealed a low authenticity of SNP sites in simplified sequencing sites, leading to difficulties in KASP primer development. Using 50 simplified sequencing varieties, KASP primers were developed using Perl programs, resulting in 28 KASP markers. Most *C. ensifolium* varieties could be easily distinguished, but there were still two groups that could not be differentiated even after multiple KASP primer developments, namely, CES001, CES002, CES019, and CES048 and CES035, and CES043. This may be due to the limited number of 28 KASP-SNP markers and the limited genomic variations that can be

revealed, necessitating the development of more *C. ensifolium* KASP-SNP markers. Additionally, the simplified genome data used for marker development cover only approximately 2% of the genome, limiting the variation that can be revealed. Furthermore, the two groups may be closely related, with minimal genetic differences (Mei, 2023). Therefore, one representative variety (CES001 and CES035) was selected from each group for fingerprint construction. Furthermore, 37 commercially cultivated *C. ensifolium* varieties were added to validate the utility of the KASP markers. These 28 markers were successfully used for genotyping 83 commercial *C. ensifolium* varieties.

The construction of a molecular fingerprint map of germplasm resources requires the use of a minimal number of primers to differentiate the maximum number of germplasms. Therefore, selecting appropriate primers is an important prerequisite for constructing a molecular fingerprint map (Wang et al., 2021b; Li et al., 2023a). Based on the principle of using the minimum number of markers to differentiate the maximum number of varieties, 11 KASP markers were selected from 28 KASP markers (PIC: 0.195-0.375; MAF: 0.125-0.479; HE: 0.219-0.499), which could efficiently differentiate all 83 *C. ensifolium* commercial varieties with at least one genotypic difference. On average, one SNP locus could identify 8 varieties, which is higher than the SNP marker differentiation efficiency developed in Chinese flowering cabbage (18 core SNP markers could completely differentiate all 89 cabbage varieties) and melon (40 core SNP markers efficiently differentiate 99% of the 259 commercial melon varieties) (Ren et al., 2023; Zhang et al., 2023a). However, it is lower than the SNP marker differentiation efficiency developed in radish, which has reached 24 accessions per SNP locus (15 core SNP markers could completely differentiate all 356 radish varieties) (Xing et al., 2024). This may be attributed to the relatively high quantity and quality of the genomic database used for radish SNP marker development (Xing et al., 2024). The development of radish markers has demonstrated the great potential and advantages of using SNP loci to differentiate varieties, prompting us to continuously expand the genomic database of *C. ensifolium* to develop more efficient SNP markers.

It is noted that there are some varieties that cannot be distinguished when using SNP loci to differentiate species, which has been observed in crops like melon and cigar tobacco (Zhang et al., 2023a; Wang et al., 2021b). In a study on cigar tobacco, Yanyan et al. utilized 47 core KASP markers to differentiate 216 cigar tobacco germplasm resources, and found that some varieties could not be distinguished. Through phenotype analysis, these were identified as synonyms (Wang et al., 2021b). The phenomenon of synonyms also exists in the *C. ensifolium* market, where vendors may alter the phenotype of orchids through physical or chemical treatments to create fake new varieties, and rename existing varieties arbitrarily (Ning, 2022; Li, 1994). This confusion in commercial varieties is detrimental to the conservation and breeding of orchids, and SNP marker-based variety identification can greatly regulate the *C. ensifolium* variety market. In cases where differentiation is not possible, DUS determination (NY/T 2441-2013 Guidelines for the conduct of tests for distinctness, uniformity and stability *Cymbidium*) can be used for further confirmation of varieties.

The PIC value is considered the most important indicator of the usefulness of molecular markers. Markers with high PIC values are usually highly polymorphic, whereas markers with low PIC values are considered less polymorphic. Furthermore, markers with PIC values >0.5 are usually considered highly polymorphic, those with PIC values of $0.25 < \text{PIC} \leq 0.5$ are usually considered medium polymorphic, and those with PIC values ≤ 0.25 are usually considered lowly polymorphic (Liu et al., 2022b). The 11 core markers required for fingerprint map construction had a PIC range of 0.195–0.375, with 90% > 0.336 indicating medium polymorphism, demonstrating the practicality and reliability of fingerprint map construction. Therefore, the 11 KASP markers are clearly reliable, effective, and accurate in detecting 83 *C. ensifolium* germplasm resources.

However, these 83 *C. ensifolium* commercial varieties constitute only a small portion of the total commercial varieties of *C. ensifolium*. To distinguish total commercial varieties, especially similar varieties, more SNP markers with greater discriminatory ability are necessary. Although only 28 out of 4,360 KASP markers were verified, verifying the remaining markers could identify more high-quality markers. These 83 commercial varieties include various typical shapes of *C. ensifolium* commercial varieties, as well as newly bred hybrid varieties in China in recent years. Additionally, utilizing SNP molecular markers and fingerprints comprehensively can effectively improve the identification capabilities of *C. ensifolium* varieties containing wild varieties, which is essential for wild resource protection, and further research is needed in this area.

Conclusion

This study provides comprehensive information about the genetic diversity of *C. ensifolium* commercial cultivars in China on the basis of a population of 50 *C. ensifolium* commercial cultivars. A series of SNPs were discovered by ddRAD-seq of 50 diverse *C. ensifolium* commercial cultivars, and the SNPs were converted into KASP panels for the genotyping of a large set of *C. ensifolium* commercial cultivars. Phylogenetic and PCA analyses revealed that the 50 *C. ensifolium* commercial cultivars were divided into four well-separated clusters, and a correlation was observed between the group distribution and the geographical origin of the *C. ensifolium* germplasm. A set of 28 KASP SNP markers was screened, and a minimum set of 11 KASP SNP markers (with 90% PIC values >0.336) was verified to distinguish 83 *C. ensifolium* commercial cultivars completely. The 11 SNP genotypes of each *C. ensifolium* variety were used to generate SNP fingerprints of a major collection (83) of cultivated *C. ensifolium* varieties in China. The KASP markers developed in this study could also be utilized for evaluating the variety authenticity of *C. ensifolium* cultivars. This is the first study to measure the diversity and population structure of a large collection of *C. ensifolium* in China on the basis of SNPs from simplified genome sequencing and the first application of KASP techniques in *C. ensifolium* for genetic studies. This is also the first study to construct a fingerprint chart of *C. ensifolium* commercial cultivars using SNP markers. The information generated in this study will aid in the selection of suitable genotypes for the breeding of new cultivars and provide a scientific basis and technical support

for the protection and identification of new *C. ensifolium* cultivars and wild germplasm resources.

Data availability statement

The datasets generated for this study can be found in the NCBI database (BioProject accession number: PRJNA1127271). 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

BS: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. AS: Data curation, Formal analysis, Methodology, Writing – review & editing. YT: Resources, Validation, Writing – review & editing. LL: Resources, Validation, Writing – review & editing. SL: Resources, Validation, Writing – review & editing. ZT: Methodology, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was supported by the Hunan Provincial Forestry Ecological Protection, Restoration and Development Special Fund Project (Xiangcai Zihuan Zhi (2022) No. 64; Xiangcai Zihuan Zhi (2023) No. 26; Xiangcai Zihuan Zhi (2023) No. 72).

Acknowledgments

We would like to thank the Department of Wildlife Conservation, The Forestry Department of Hunan Province, for providing project financial support and the staff of Nanjing Genepioneer Biotechnologies Co., Ltd., for their support in simplified genome sequencing and SNP analyses and verification. We also thank the peer reviewers for their suggestions.

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.2024.1460603/full#supplementary-material>

SUPPLEMENTARY EXCEL 1

Information on 28 *C. ensifolium* KASP marker primers.

SUPPLEMENTARY EXCEL 2

Information on 4360 *C. ensifolium* SNP loci.

SUPPLEMENTARY EXCEL 3

Statistical table of the number of variations in the SNP type.

SUPPLEMENTARY EXCEL 4

Genetic diversity information of 28 KASP markers in *C. ensifolium* materials.

SUPPLEMENTARY EXCEL 5

KASP genotyping results of 28 markers for 83 *C. ensifolium* varieties.

References

- Addison, C. K., Angira, B., Kongchum, M., Harrell, D. L., Baisakh, N., Linscombe, S. D., et al. (2020). Characterization of haplotype diversity in the BADH2 aroma Gene and development of a KASP SNP assay for predicting aroma in U.S. Rice. *Rice (N Y)* 13, 47. doi: 10.1186/s12284-020-00410-7
- Ai, Y., Chen, L., Xie, T. X., Chen, J., Lan, S. R., and Peng, D. H. (2019). Construction of core collection of *Cymbidium ensifolium* cultivars based on SSR fluorescent markers. *Acta Hortic. Sin.* 46, 1999–2008. doi: 10.16420/j.issn.0513-353x.2019-0327
- Ai, Y., Li, Z., Sun, W.-H., Chen, J., Zhang, D., Ma, L., et al. (2021). The *Cymbidium* genome reveals the evolution of unique morphological traits. *Horticulture Res.* 8, 255. doi: 10.1038/s41438-021-00683-z
- Ai, Y., Zheng, Q. D., Wang, M. J., Xiong, L. W., Li, P., Guo, L. T., et al. (2023). Molecular mechanism of different flower color formation of *Cymbidium ensifolium*. *Plant Mol. Biol.* 113, 193–204. doi: 10.1007/s11033-023-01382-0
- Akpertey, A., Padi, F. K., Meinhardt, L., and Zhang, D. (2020). Effectiveness of single nucleotide polymorphism markers in genotyping germplasm collections of *Coffea canephora* using KASP assay. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.612593
- Amanullah, S., Osae, B. A., Yang, T., Li, S., Abbas, F., Liu, S., et al. (2022). Development of whole genome SNP-CAPS markers and preliminary QTL mapping of fruit pedicel traits in watermelon. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.879919
- Ashwath, M. N., Lavale, S. A., Santhoshkumar, A. V., Mohapatra, S. R., Bhardwaj, A., Dash, U., et al. (2023). Genome-wide association studies: an intuitive solution for SNP identification and gene mapping in trees. *Funct. Integr. Genomics* 23, 297. doi: 10.1007/s10142-023-01224-8
- Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., et al. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PLoS One* 14, e0217222. doi: 10.1371/journal.pone.0217222
- 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
- Button, P. (2008). The international union for the protection of new varieties of plants (UPOV) recommendations on variety denominations. *Acta Hortic.* 799, 191–200. doi: 10.17660/ActaHortic.2008.799.27
- Cao, H., Li, H., Chen, X., Zhang, Y., Lu, L., Li, S., et al. (2022). Insight into the molecular mechanisms of leaf coloration in *Cymbidium ensifolium*. *Front. Genet.* 13. doi: 10.3389/fgene.2022.923082
- Chen, M. K., Chen, L., Sun, W. H., Ma, S. H., Lan, S. R., Peng, D. H., et al. (2022). Genetic diversity analysis and core collection of *Cymbidium ensifolium* germplasm resources. *Acta Hortic. Sin.* 49, 175–186. doi: 10.16420/j.issn.0513-353x.2021-0312
- Chen, Y., Tang, C. L., Li, H., Qian, L., Luo, S. K., and Zhou, X. R. (2020). Analysis of affinity of *Bauhinia variegata* germplasm resources based on ISSR markers. *Southwest China J. Agric. Sci.* 33, 26–31. doi: 10.16213/j.cnki.scjas.2020.1.005
- Chen, Z., Tang, D., Ni, J., Li, P., Wang, L., Zhou, J., et al. (2021). Development of genic KASP SNP markers from RNA-Seq data for map-based cloning and marker-assisted selection in maize. *BMC Plant Biol.* 21, 157. doi: 10.1186/s12870-021-02932-8
- Chen, C., Wu, Y., Li, J., Wang, X., Zeng, Z., Xu, J., et al. (2023). TBtools-II: A “one for all, all for one” bioinformatics platform for biological big-data mining. *Mol. Plant* 16, 1733–1742. doi: 10.1016/j.molp.2023.09.010
- Clare, S. J., King, R. M., Hardigan, M., Dossett, M., Montanari, S., Chagné, D., et al. (2023). Development of KASP fingerprinting panel for clonal identification in red raspberry (*Rubus idaeus* L.). *Plant Breed.* 142, 798–808. doi: 10.1111/pbr.13141
- Congiu, L., Chicca, M., Cella, R., Rossi, R., and Bernacchia, G. (2000). The use of random amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application. *Mol. Ecol.* 9, 229–232. doi: 10.1046/j.1365-294x.2000.00811.x
- Cui, X. Q., Huang, C. Y., Deng, J. L., Li, X. M., Li, X. L., and Zhang, Z. B. (2023). SNP markers development and genetic relationship analysis of *Dendrobium* germplasms using SLAF-seq technology. *Biotechnol. Bull.* 39, 141–148. doi: 10.13560/j.cnki.biotech.bull.1985.2022-1386
- Dipta, B., Sood, S., Mangal, V., Bhardwaj, V., Thakur, A. K., Kumar, V., et al. (2024). KASP: a high-throughput genotyping system and its applications in major crop plants for biotic and abiotic stress tolerance. *Mol. Biol. Rep.* 51, 508. doi: 10.1007/s11033-024-09455-z
- Fleming, M. B., Miller, T., Fu, W., Li, Z., Gasic, K., and Saski, C. (2022). Ppe.XapF: High throughput KASP assays to identify fruit response to *Xanthomonas arboricola* pv. pruni (Xap) in peach. *PLoS One* 17, e0264543. doi: 10.1371/journal.pone.0264543
- Franklin, R. N., Karim, N., Goecker, Z. C., Durbin-Johnson, B. P., Rice, R. H., and Parker, G. J. (2020). Proteomic genotyping: Using mass spectrometry to infer SNP genotypes in pigmented and non-pigmented hair. *Forensic Sci. Int.* 310, 110200. doi: 10.1016/j.forsciint.2020.110200
- Gazendam, L., Mojapelo, P., and Bairu, M. W. (2022). Potato cultivar identification in south Africa using a custom SNP panel. *Plants (Basel)* 11, 1546. doi: 10.3390/plants11121546
- Geng, X., Qu, Y., Jia, Y., He, S., Pan, Z., Wang, L., et al. (2021). Assessment of heterosis based on parental genetic distance estimated with SSR and SNP markers in upland cotton (*Gossypium hirsutum* L.). *BMC Genomics* 22, 123. doi: 10.1186/s12864-021-07431-6
- Gerasimova, Y. V., Ballantyne, J., and Kolpashchikov, D. M. (2013). Detection of SNP-containing human DNA sequences using a split sensor with a universal molecular beacon reporter. *Methods Mol. Biol.* 1039, 69–80. doi: 10.1007/978-1-62703-535-4_5
- Gomes, S., Castro, C., Barrias, S., Pereira, L., Jorge, P., Fernandes, J. R., et al. (2018). Alternative SNP detection platforms, HRM and biosensors, for varietal identification in *Vitis vinifera* L. using F3H and LDOX genes. *Sci. Rep.* 8, 5850. doi: 10.1038/s41598-018-24158-9
- Guan, S., Li, W., Jin, H., Zhang, L., and Liu, G. (2023). Development and validation of a 54K genome-wide liquid SNP chip panel by target sequencing for dairy goat. *Genes (Basel)* 14, 1122. doi: 10.3390/genes14051122
- He, C., Holme, J., and Anthony, J. (2014). SNP genotyping: the KASP assay. *Methods Mol. Biol.* 1145, 75–86. doi: 10.1007/978-1-4939-0446-4_7
- Hong, S., Choi, S. R., Kim, J., Jeong, Y. M., Kim, J. S., Ahn, C. H., et al. (2022). Identification of accession-specific variants and development of KASP markers for assessing the genetic makeup of *Brassica rapa* seeds. *BMC Genomics* 23, 326. doi: 10.1186/s12864-022-08567-9
- Hu, W., Huang, R. Z., Pan, X. H., Li, J. F., and Sun, D. (2008). RAPD analysis of thirty-eight *Cymbidium ensifolium* cultivars. *Acta Hortic. Sin.* 2, 289–294. doi: 10.16420/j.issn.0513-353x.2008.02.012
- Huang, H. X. (2012). *Study on the economic and the cultural value in Chinese orchids* (Fujian: Fujian Agriculture and Forestry University) 2012.
- Jeffreys, A. J. (2013). The man behind the DNA fingerprints: an interview with Professor Sir Alec Jeffreys. *Invest. Genet.* 4, 21. doi: 10.1186/2041-2223-4-21
- Jimoh, T. O., Costa, B. C., Chansrinoyom, C., Chaotham, C., Chanvorachote, P., Rojsittisak, P., et al. (2022). Three new dihydrophenanthrene derivatives from *Cymbidium ensifolium* and their cytotoxicity against cancer cells. *Molecules* 27, 2222. doi: 10.3390/molecules27072222
- Jin, F. (2019). *Analysis of Cymbidium ensifolium transcriptomic database, and the development and application of Genic-SSR markers* (Hubei: Hubei University) 2019.
- Josia, C., Mashingaidze, K., Amelework, A. B., Kondwakwenda, A., Musvosvi, C., and Sibiyi, J. (2021). SNP-based assessment of genetic purity and diversity in maize hybrid breeding. *PLoS One* 16, e0249505. doi: 10.1371/journal.pone.0249505
- Kovalchuk, S. N., and Arkhipova, A. L. (2023). Development of TaqMan PCR assay for genotyping SNP rs211250281 of the bovine agpat6 gene. *Anim. Biotechnol.* 34, 3250–3255. doi: 10.1080/10495398.2022.2077742

- Kuang, M., Wei, S. J., Wang, Y. Q., Zhou, D. Y., Ma, L., Fang, D. D., et al. (2016). Development of a core set of SNP markers for the identification of upland cotton cultivars in China. *J. Integr. Agr.* 15, 954–962. doi: 10.1016/S2095-3119(16)61226-6
- Kumar, S., Pradhan, A. K., Kumar, U., Dhillon, G. S., Kaur, S., Budhlakoti, N., et al. (2022). Validation of Novel spot blotch disease resistance alleles identified in unexplored wheat (*Triticum aestivum* L.) germplasm lines through KASP markers. *BMC Plant Biol.* 22, 618. doi: 10.1186/s12870-022-04013-w
- Li, X. H. (1994). How to identify fake “precious orchids. *China Flower Penjing* 10, 32. Available at: https://kns.cnki.net/kcms2/article/abstract?v=Mw9fkKjKlJppdeRanRxBKkpcuurTbTis0ud67HjxjK7GQKMEZN6JssxD55mENX3vaGslc0b9n2q2EMw61oQFnu7EhvyOy0rdaiy2dKbF4dTTSRHMg62A2Nu2S-oMDMVkp_ss1hrpuEi0f-lkSIV7W1jyTcfbrq6EHIt4LqkSfhkjj4U3XqMTMyS41zO2NvU&uniplatform=NZKPT&language=CHS
- Li, Y. (2014). *Population genetic structure of Cymbidium ensifolium from Jianxi based on ITS and cpDNA molecular markers* (Nanchang: Nanchang University) 2014.
- Li, J., Chang, X., Huang, Q., Liu, P., Zhao, X., Li, F., et al. (2023a). Construction of SNP fingerprint and population genetic analysis of honeysuckle germplasm resources in China. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1080691
- 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, Y. G., Guo, W. H., and Wu, B. J. (2002). Studies on karyotypes of four species of *Cymbidium* in China. *Acta Botanica Boreali-Occidentalia Sin.* 22, 1438–1444. Available at: https://kns.cnki.net/kcms2/article/abstract?v=Mw9fkKjKlJqfpZL1mxEFSS5v5nXFXt3f9KXcKGmaNKKW45jj3fDke0EyYqt3adfypl1uwellenSSXJkpcmfWXXR2sN5rFr-PBXOgZtcU2tdoqRZGZJj3GS9X0lyguQGfvel7q2SG9-okgrVlfHVsaAgO_2fv8qi7Zn2uisEgqf-lidS8pd2_zZeHouFnliul&uniplatform=NZKPT&language=CHS
- 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, X., Jin, F., Jin, L., Jackson, A., Huang, C., Li, K., et al. (2014). Development of *Cymbidium ensifolium* genic-SSR markers and their utility in genetic diversity and population structure analysis in cymbidiums. *BMC Genet.* 15, 124. doi: 10.1186/s12863-014-0124-5
- Li, X., Liu, L., Sun, S., Li, Y., Jia, L., Ye, S., et al. (2023b). Transcriptome analysis reveals the key pathways and candidate genes involved in salt stress responses in *Cymbidium ensifolium* leaves. *BMC Plant Biol.* 23, 64. doi: 10.1186/s12870-023-04050-z
- Li, Z., Yu, H., Li, X., Zhang, B., Ren, W., Liu, X., et al. (2020). Competitive allele-specific PCR (KASP) genotyping and heterotic group classification of 244 inbred lines in cabbage (*Brassica oleracea* L. var. capitata). *Euphytica* 216, 106. doi: 10.1007/s10681-020-02640-8
- Lin, L., Wang, F., Wu, M., and Wang, S. (2022). ddRAD sequencing-based scanning of genetic variants in *Sargassum fusiforme*. *J. Mar. Sci. Eng.* 10, 958. doi: 10.3390/jmse10070958
- Liu, X. F., Li, X. M., Zhang, G. W., Zhang, P. J., Liu, G., Huang, M., et al. (2022a). Population genetic analysis of *Jacaranda mimosifolia* by RAD high throughput sequencing technique. *J. Trop. Subtrop Bot.* 30, 613–622. doi: 10.11926/jtsb.4517
- Liu, G., Liu, D., Zhang, A., Liu, H., Mía, M. S., Mullan, D., et al. (2023). Identification of KASP markers and candidate genes for drought tolerance in wheat using 90K SNP array genotyping of near-isogenic lines targeting a 4BS quantitative trait locus. *Theor. Appl. Genet.* 136, 190. doi: 10.1007/s00122-023-04438-3
- Liu, Y., Zheng, Y. Q., Li, C. H., Lin, F. R., and Huang, P. (2022b). Genomic characteristics and population genetic variation of *Dalbergia cultrata* Graham ex Benth in China. *For. Res.* 35, 44–53. doi: 10.13275/j.cnki.lykxyj.2022.004.005
- Magbanua, Z. V., Hsu, C. Y., Pechanova, O., Arick, M. 2nd, Grover, C. E., and Peterson, D. G. (2023). Innovations in double digest restriction-site associated DNA sequencing (ddRAD-Seq) method for more efficient SNP identification. *Anal. Biochem.* 662, 115001. doi: 10.1016/j.ab.2022.115001
- Mannino, G., Kunz, R., and Maffei, M. E. (2023). Discrimination of green coffee (Coffea arabica and Coffea canephora) of different geographical origin based on antioxidant activity, high-throughput metabolomics, and DNA RFLP fingerprinting. *Antioxidants (Basel)* 12, 1135. doi: 10.3390/antiox12051135
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Mei, X. (2023). *Identification of cowpea germplasm resources by KASP-SNP molecular markers and screening of cowpea regeneration genotypes* (Wuhan: Jiangnan University) 2023.
- Mei, H., Zhang, X., Zhao, F., Ruan, R., and Fu, Q. (2024). Integrated metabolome and transcriptome analysis provides insight into the leaf color change of *Cymbidium ensifolium*. *Acta Physiologiae Plantarum* 46, 50. doi: 10.1007/s11738-024-03671-7
- Morales, K. Y., Singh, N., Perez, F. A., Ignacio, J. C., Thapa, R., Arbelaez, J. D., et al. (2020). An improved 7K SNP array, the C7AIR, provides a wealth of validated SNP markers for rice breeding and genetics studies. *PLoS One* 15, e0232479. doi: 10.1371/journal.pone.0232479
- Nguyen, N. N., Kim, M., Jung, J. K., Shim, E. J., Chung, S. M., Park, Y., et al. (2020). Genome-wide SNP discovery and core marker sets for assessment of genetic variations in cultivated pumpkin (*Cucurbita* spp.). *Hortic. Res.* 7, 121. doi: 10.1038/s41438-020-00342-9
- Ning, H. J. (2022). *Classification of species and varieties of Chinese orchid* (Beijing: Beijing Forestry University) 2022.
- Patterson, E. L., Fleming, M. B., Kessler, K. C., Nissen, S. J., and Gaines, T. A. (2017). A KASP genotyping method to identify northern watermilfoil, eurasian watermilfoil, and their interspecific hybrids. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00752
- 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
- Pradhan, A. K., Budhlakoti, N., Chandra Mishra, D., Prasad, P., Bhardwaj, S. C., Sareen, S., et al. (2023). Identification of novel QTLs/Defense genes in spring wheat germplasm panel for seedling and adult plant resistance to stem rust and their validation through KASP marker assays. *Plant Dis.* 107, 1847–1860. doi: 10.1094/PDIS-09-22-2242-RE
- Ramesh, P., Mallikarjuna, G., Sameena, S., Kumar, A., Gurulakshmi, K., Reddy, B. V., et al. (2020). Advancements in molecular marker technologies and their applications in diversity studies. *J. Biosci.* 45, 123. doi: 10.1007/s12038-020-00089-4
- Ren, H. L., Xu, D. L., Zhang, J., Zou, J. W., Li, G. G., Zhou, X. Y., et al. (2023). Establishment of SNP fingerprinting and identification of Chinese flowering cabbage varieties based on KASP genotyping. *Acta Hort.* 50, 307–318. doi: 10.16420/j.issn.0513-353x.2021-1046
- Roncallo, P. F., Beaufort, V., Larsen, A. O., Dreisigacker, S., and Echenique, V. (2019). 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, e0218562. doi: 10.1371/journal.pone.0218562
- Shen, Y., Wang, J., Shaw, R. K., Yu, H., Sheng, X., Zhao, Z., et al. (2021). Development of GBTS and KASP panels for genetic diversity, population structure, and fingerprinting of a large collection of broccoli (*Brassica oleracea* L. var. italica) in China. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.655254
- Sivu, A. R., Pradeep, N. S., Pandurangan, A. G., Dwivedi, M. D., and Pandey, A. K. (2022). Phylogenetic relationships of Indian *Memecylon* L. (Melastomataceae) based on nrDNA ITS and cpDNA rbcL sequence data. *J. Genet.* 101, 8. doi: 10.1007/s12041-021-01343-w
- Su, J., Zhang, F., Chong, X., Song, A., Guan, Z., Fang, W., et al. (2019). Genome-wide association study identifies favorable SNP alleles and candidate genes for waterlogging tolerance in chrysanthemums. *Hortic. Res.* 6, 21. doi: 10.1038/s41438-018-0101-7
- Sun, N., Chen, J., Wang, Y., Hussain, I., Lei, N., Ma, X., et al. (2023). Development and utility of SSR markers based on Brassica sp. whole-genome in triangle of U. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1259736
- Tang, W., Lin, J., Wang, Y., An, H., Chen, H., Pan, G., et al. (2022). Selection and validation of 48 KASP markers for variety identification and breeding guidance in conventional and hybrid rice (*Oryza sativa* L.). *Rice (N Y)* 15, 48. doi: 10.1186/s12284-022-00594-0
- Tian, H., Yang, Y., Wang, R., Fan, Y., Yi, H., Jiang, B., et al. (2021). Screening of 200 core SNPs and the construction of a systematic SNP-DNA standard fingerprint database with more than 20,000 maize varieties. *Agriculture* 11, 597. doi: 10.3390/agriculture11070597
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001
- Wang, H. L. (2020). *Evaluation of phenotypic traits and genetic diversity of 39 Cymbidium ensifolium* (Guangxi: Guangxi University) 2020.
- Wang, H. L., Bu, Z. Y., Zeng, Y. H., and Long, Q. Y. (2021a). Study on genetic diversity of *Cymbidium ensifolium* germplasm based on ISSR marker. *Chin. J. Trop. Crops* 42, 2526–2534. doi: 10.3969/j.issn.1000-2561.2021.09.012
- Wang, F. Q., Fan, X. C., Zhang, Y., Liu, C. H., and Jiang, J. F. (2020). Application and prospect of SNP molecular markers in crop variety identification. *J. Plant Genet. Resour* 21, 1308–1320. doi: 10.13430/j.cnki.jpgr.20200309002
- Wang, F. Q., Fan, X. C., Zhang, Y., Sun, L., Liu, C. H., and Jiang, J. F. (2022). Establishment and application of an SNP molecular identification system for grape cultivars. *J. Integr. Agr.* 21, 1044–1057. doi: 10.1016/S2095-3119(21)63654-7
- Wang, Q., Fu, L., Zhang, X., Dai, X., Bai, M., Fu, G., et al. (2016). Expansion of a SNaPshot assay to a 55-SNP multiplex: Assay enhancements, validation, and power in forensic science. *Electrophoresis* 37, 1310–1317. doi: 10.1002/elps.201500353
- Wang, H.-Z., Lu, J.-J., Hu, X., and Liu, J.-J. (2011). Genetic variation and cultivar identification in *Cymbidium ensifolium*. *Plant Systematics Evol.* 293, 101–110. doi: 10.1007/s00606-011-0429-z
- Wang, Y., Lv, H., Xiang, X., Yang, A., Feng, Q., Dai, P., et al. (2021b). Construction of a SNP fingerprinting database and population genetic analysis of cigar tobacco germplasm resources in China. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.618133
- Wang, Y., Shahid, M. Q., Ghouri, F., Ercisli, S., and Baloch, F. S. (2019). Development of EST-based SSR and SNP markers in *Gastrodia elata* (herbal medicine) by sequencing, *de novo* assembly and annotation of the transcriptome. *3 Biotech.* 9, 292. doi: 10.1007/s13205-019-1823-4
- Wang, B., Tan, H. W., Fang, W., Meinhardt, L. W., Mischke, S., Matsumoto, T., et al. (2015). Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (*Dimocarpus longan*) germplasm. *Hortic. Res.* 2, 14065. doi: 10.1038/hortres.2014.65

- Wang, X. Y., and Wang, C. X. (2014). Genetic diversity analysis of *Cymbidium* cultivars by AFLP molecular markers. *J. Agric.* 4, 69–72. Available at: https://kns.cnki.net/kcms2/article/abstract?v=Mw9fkKjKljqOFIqglCf3HBLbbxm3TpLcOazXokW7QsUrVx_lh9ZUvdPhS7seSq-KCUDfJP1ox8PdSl-c7SoZMUueRwQrSWGkedEwt74R6xvuiq7eEnPSpbmASgn-ZGd5BQd6Kx4zgkcPjTpsvY2CYjo1_PCtsQ-hDCgEXBk7ltt5V-dihdML_oPkv0sDoxM4&uniplatform=NZKPT&language=CHS
- Wei, X. Y., Liu, H., Ma, H., Bie, T. D., and Sun, Y. (2024). Analysis of genetic diversity and fingerprint construction of 96 *Cymbidium* based on ISSR markers. *J. Plant Genet. Resour* 25, 586–599. doi: 10.13430/j.cnki.jpgr.20230810003
- Winfield, M., Burridge, A., Ordidge, M., Harper, H., Wilkinson, P., Thorogood, D., et al. (2020). Development of a minimal KASP marker panel for distinguishing genotypes in apple collections. *PLoS One* 15, e0242940. doi: 10.1371/journal.pone.0242940
- Xing, X., Hu, T., Wang, Y., Li, Y., Wang, W., Hu, H., et al. (2024). Construction of SNP fingerprints and genetic diversity analysis of radish (*Raphanus sativus* L.). *Front. Plant Sci.* 15. doi: 10.3389/fpls.2024.1329890
- Yang, G., Chen, S., Chen, L., Sun, K., Huang, C., Zhou, D., et al. (2019). Development of a core SNP arrays based on the KASP method for molecular breeding of rice. *Rice (N Y)* 12, 21. doi: 10.1186/s12284-019-0272-3
- Yang, F., Guo, Y., Li, J., Lu, C., Wei, Y., Gao, J., et al. (2023). Genome-wide association analysis identified molecular markers and candidate genes for flower traits in Chinese orchid (*Cymbidium sinense*). *Hortic. Res.* 10, uhad206. doi: 10.1093/hr/uhad206
- Yang, Y., Lyu, M., Liu, J., Wu, J., Wang, Q., Xie, T., et al. (2022). Construction of an SNP fingerprinting database and population genetic analysis of 329 cauliflower cultivars. *BMC Plant Biol.* 22, 522. doi: 10.1186/s12870-022-03920-2
- Yang, X. F., Zhang, Z. Y., Miao, L. X., Zhang, Y. C., Shen, L., Qin, Q. P., et al. (2020). Preliminary analysis on the molecular differences of fruit colouring between Yuexin strawberry and its somaclonal mutation. *Acta Hortic. Sin.* 47, 1999–2008. doi: 10.16420/j.issn.0513-353x.2019-0996
- Yang, F., and Zhu, G. (2015). Digital gene expression analysis based on *De Novo* transcriptome assembly reveals new genes associated with floral organ differentiation of the orchid plant *Cymbidium ensifolium*. *PLoS One* 10, e0142434. doi: 10.1371/journal.pone.0142434
- Yuan, J., Wen, Z., Gu, C., and Wang, D. (2014). Introduction of high throughput and cost effective SNP genotyping platforms in Soybean. *Plant Genetics Genomics Biotechnol.* 2, 90–94. doi: 10.5147/pggb.2014.0144
- Zeng, Z., Guo, C., Yan, X., Song, J., Wang, C., Xu, X., et al. (2022). QTL mapping and KASP marker development for seed vigor related traits in common wheat. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.994973
- Zhang, Z., Cao, Y., Wang, Y., Yu, H., Wu, H., Liu, J., et al. (2023b). Development and validation of KASP markers for resistance to *Phytophthora capsici* in *Capsicum annuum* L. *Mol. Breed* 43, 20. doi: 10.1007/s11032-023-01367-3
- Zhang, S., Li, Y., Li, Y., Zhang, Y., Hao, Y., Hou, Z., et al. (2024). Development of SSR markers for and fingerprinting of walnut genetic resources. *Forests* 15, 405. doi: 10.3390/f15030405
- Zhang, Y., Wang, J., Yang, L., Niu, J., Huang, R., Yuan, F., et al. (2022b). Development of SSR and SNP markers for identifying opium poppy. *Int. J. Legal Med.* 136, 1261–1271. doi: 10.1007/s00414-022-02810-4
- Zhang, J., Yang, J., Fu, S., Ren, J., Zhang, X., Xia, C., et al. (2022a). Comparison of DUS testing and SNP fingerprinting for variety identification in cucumber. *Hortic. Plant J.* 8, 575–582. doi: 10.1016/j.hpi.2022.07.002
- Zhang, J., Yang, J., Lv, Y., Zhang, X., Xia, C., Zhao, H., et al. (2023a). Genetic diversity analysis and variety identification using SSR and SNP markers in melon. *BMC Plant Biol.* 23, 39. doi: 10.1186/s12870-023-04056-7
- Zhang, J., Yang, J., Zhang, L., Luo, J., Zhao, H., Zhang, J., et al. (2020). A new SNP genotyping technology Target SNP-seq and its application in genetic analysis of cucumber varieties. *Sci. Rep.* 10, 5623. doi: 10.1038/s41598-020-62518-6
- Zhao, Y., Chen, W., Cui, Y., Sang, X., Lu, J., Jing, H., et al. (2021). Detection of candidate genes and development of KASP markers for Verticillium wilt resistance by combining genome-wide association study, QTL-seq and transcriptome sequencing in cotton. *Theor. Appl. Genet.* 134, 1063–1081. doi: 10.1007/s00122-020-03752-4
- Zhao, Y., Liu, X. D., Zhao, H. K., Yuan, C. P., Qi, G. X., Wang, Y. M., et al. (2017). Comparison of methods for SNP genotyping in soybean. *Mol. Plant Breed* 15, 3540–3546. doi: 10.13271/j.mpb.015.003540
- Zongze, Z., Xin, Y., Awais, A. A., Weiqiang, L., Chunqun, W., Di, W., et al. (2018). Development of a tetra-primer ARMS-PCR for detecting the E198A SNP in the isotype-1 beta-tubulin gene of *Haemonchus contortus* populations in China. *Vet. Parasitol.* 252, 127–130. doi: 10.1016/j.vetpar.2018.01.021



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Dharmaprakash Pandhari Wankhede,
Indian Council of Agricultural Research
(ICAR), India

*CORRESPONDENCE

Daiana A. Zhernova

✉ zhernova.d@ya.ru

Nataliya V. Melnikova

✉ mnv-4529264@yandex.ru

RECEIVED 12 September 2024

ACCEPTED 10 December 2024

PUBLISHED 15 January 2025

CITATION

Zhernova DA, Pushkova EN, Rozhmina TA,
Borkhert EV, Arkhipov AA, Sigova EA,
Dvorianinova EM, Dmitriev AA and
Melnikova NV (2025) History and prospects
of flax genetic markers.

Front. Plant Sci. 15:1495069.

doi: 10.3389/fpls.2024.1495069

COPYRIGHT

© 2025 Zhernova, Pushkova, Rozhmina,
Borkhert, Arkhipov, Sigova, Dvorianinova,
Dmitriev and Melnikova. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

History and prospects of flax genetic markers

Daiana A. Zhernova^{1*}, Elena N. Pushkova¹,
Tatiana A. Rozhmina², Elena V. Borkhert¹,
Alexander A. Arkhipov^{1,3}, Elizaveta A. Sigova^{1,4},
Ekaterina M. Dvorianinova^{1,4}, Alexey A. Dmitriev^{1,4}
and Nataliya V. Melnikova^{1*}

¹Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, ²Federal Research Center for Bast Fiber Crops, Torzhok, Russia, ³I.M. Sechenov First Moscow State Medical University, Moscow, Russia, ⁴Moscow Institute of Physics and Technology, Moscow, Russia

Flax (*Linum usitatissimum* L.) is known as a dual-purpose crop, producing both fiber and oil, which have a wide range of uses. Successful flax breeding requires knowledge on the genetic determinants of flax traits. The former identification of molecular markers for valuable traits used labor-intensive and sometimes poorly reproducible approaches. However, they allowed an assessment of the genetic diversity of flax and its relatives, the construction of linkage maps, and the identification of some markers for important characteristics. The sequencing of flax whole genome triggered the development of genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping. QTLs and quantitative trait nucleotides (QTNs) were identified for valuable seed- and fiber-related features and for resistance to biotic and abiotic stressors. Cost-effective and accurate analysis of large number of genotypes for multiple markers simultaneously using microarrays or targeted deep sequencing became available, as well as HRM, TaqMan, KASP, and other fluorescence-based high-throughput methods for detecting DNA polymorphisms. However, most DNA markers identified in flax are ambiguously linked to trait expression and are not universally applicable. A major challenge remains the lack of knowledge on functional polymorphisms. To date, only a few are known, mainly mutations in the *FAD3* genes responsible for reduced linolenic acid content in linseed oil. For the further development of marker-assisted and genomic selection of flax, it is necessary to analyze exhaustively phenotyped sample sets, to identify DNA polymorphisms that determine valuable traits, and to develop efficient DNA test systems.

KEYWORDS

flax, *Linum usitatissimum* L., DNA polymorphisms, DNA markers, GWAS, QTL, QTN, marker-assisted selection

1 Introduction

Flax (*Linum usitatissimum* L.) is one of the world's ancient crops (family Linaceae DC. ex Perleb), known since ancient Egypt and Mesopotamia (Saha and Hazra, 2004), although it was domesticated only 8,000–10,000 years ago. Flax is a dicotyledonous self-pollinated annual herbaceous plant (diploid chromosome set $2n = 2x = 30$, genome size ~ 450 Mb (Green et al., 2008; Soto-Cerda et al., 2013; Dvorianinova et al., 2022; You et al., 2023)). Flax is known as a dual-purpose crop, producing both fiber and oil. *L. usitatissimum* is one of the three most important textile crops in the world and one of the five most important oilseed crops (Deng et al., 2011; Ottai et al., 2011). Flax seed is a rich source of ω -3 (linolenic) and ω -6 (linoleic) fatty acids, easily digestible dietary fiber, and lignans, which have beneficial effects on the human body. In particular, they prevent and reduce the risk of cardiovascular diseases, various types of cancer, diabetes, atherosclerosis, arthritis, osteoporosis, autoimmune diseases, reduce cholesterol levels, stimulate blood flow, have a beneficial effect on the nervous system, and have antioxidant, anti-inflammatory, and antimicrobial effects (Kezimana et al., 2018; Campos et al., 2019; Parikh et al., 2019; Saini et al., 2021; Al-Madhagy et al., 2023). In addition, flax seeds are widely used in pharmaceuticals, animal feed, and for the production of environmentally friendly paints, varnishes, lubricants, linoleum, biofuels, etc (Singh et al., 2011; Corino et al., 2014; Goyal et al., 2014; Campos et al., 2019). Another valuable product derived from flax is fiber, which is used in the textile industry and for the production of paper and a variety of composite materials (Jhala and Hall, 2010; Fombuena et al., 2019; Asyraf et al., 2022).

The aim of our review is to describe the history of development of flax DNA markers and to propose the promising directions of their future in flax breeding for creation of varieties with a complex of valuable traits.

2 Development and application of molecular markers in flax

The beginning of the new millennium was characterized by a great interest in the use of molecular markers for plant studies. The first studies on flax used mainly dominant (Random Amplified Polymorphic DNA – RAPD, Amplified Fragment Length Polymorphism – AFLP, Inter Simple Sequence Repeat – ISSR, Inter-Retrotransposon Amplified Polymorphism – IRAP, Retrotransposon Microsatellite Amplified Polymorphism – REMAP, Sequence-Specific Amplified Polymorphism – SSAP, and Inter-Primer Binding Site – iPBS) and codominant (Restriction Fragment Length Polymorphism – RFLP and microsatellites or Simple Sequence Repeats – SSRs) markers.

RFLP markers were used in studies of flax varieties and hybrids to identify molecular markers for valuable traits and to characterize genetic diversity (Schneeberger and Cullis, 1991; Lawrence et al., 1993; Oh et al., 2000).

RAPD method was applied in studies of flax varieties, hybrids, landraces, and related to *L. usitatissimum* species for genotyping, characterization of genetic diversity, determination of phylogenetic relationships, and identification of molecular markers for valuable traits (Cullis et al., 1999; Stegniń et al., 2000; Bo et al., 2002; Fu et al., 2002a, 2002b, 2003a, 2003b; Muravenko et al., 2003; Fu, 2005; Diederichsen and Fu, 2006; Bo et al., 2008; Muravenko et al., 2009; Singh et al., 2009; Bibi et al., 2015; Nagabhushanam et al., 2021).

AFLP approach also enabled genotyping flax varieties, analyzing their genetic diversity, assessing genetic relationships between genotypes, and identifying markers for valuable flax traits using mapping populations (Spielmeyer et al., 1998; Everaert et al., 2001; Van Treuren et al., 2001; Wakjira et al., 2005; Adugna et al., 2006; Chandrawati et al., 2014).

SSR markers were widely used in flax studies to genotype varieties, evaluate their genetic diversity, assess their phylogenetic relationships, create their genetic maps, and identify markers for valuable traits using mapping populations and accession collections (Ellegren, 2004; Rashid and Duguid, 2005; Roose-Amsaleg et al., 2006; Cloutier et al., 2009; Fu and Peterson, 2010; Deng et al., 2011; Rachinskaya et al., 2011; Soto-Cerda et al., 2011; Cloutier et al., 2012; Soto-Cerda et al., 2012; Asgarinia et al., 2013; Soto-Cerda et al., 2013, 2014; Chandrawati et al., 2017b; Choudhary et al., 2017a, 2017; Sudarshan et al., 2017; Wu et al., 2017; Chandrawati et al., 2017a; Saha et al., 2019; Soto-Cerda et al., 2019; Singh et al., 2021; Saroha et al., 2022b; Chen and Liu, 2024).

ISSR markers were applied to characterize the genetic diversity and relationships among flax varieties and *L. usitatissimum* wild ancestor *Linum bienne* Mill. and to identify markers for valuable traits (Wiesnerova and Wiesner, 2004; Rajwade et al., 2010; Uysal et al., 2010; El Sayed et al., 2018; Kocak et al., 2023).

Markers based on the analysis of sequences derived from retrotransposons were also used in flax research. SSAP, IRAP, REMAP, and iPBS markers allowed evaluation of genetic diversity and genotyping of flax varieties, characterization of relationships between species of the genus *Linum*, and assessment of activity of LTR retrotransposons in flax plants exposed to stressors (Smýkal et al., 2011; Melnikova et al., 2014; Habibollahi et al., 2015; Abbasi Holasou et al., 2016; Lancíková and Žiarovská, 2020; Žiarovská et al., 2022).

The combination of different types of dominant and codominant markers was utilized in flax studies for linkage map development, genotyping, and evaluation of genetic diversity and relationships among varieties (Oh et al., 2000; Kumari et al., 2018; Mhired and Heslop-Harrison, 2018; Osman et al., 2021).

The molecular markers described above have significant weaknesses: insufficient number in the genome, inapplicability for complex polygenic traits, poor reproducibility, and labor-intensive experiments. The period of active use of dominant and codominant markers was replaced by the era of Single Nucleotide Polymorphisms (SNPs), which became possible due to advances in DNA sequencing technologies. The phylogenetic relationships of species in the genus *Linum* were studied using sequencing of non-

coding regions of chloroplast DNA, and some hypotheses on the origin of cultivated flax were confirmed (Fu and Allaby, 2010). The combined use of EST (expressed sequence tag)-SSR and SNP markers on a doubled haploid population allowed the development of a linkage map for flax and the localization of markers for seed traits on chromosomes (Cloutier et al., 2011). However, a new level of flax research was opened by the transition from sequencing single loci to sequencing whole genomes or large sets of genomic regions. The sequencing of about 44 thousand BAC clones of the variety CDC Bethune enabled the construction of a genome-wide physical map of flax and the characterization of its genome (Ragupathy et al., 2011). The further development of flax molecular markers was largely driven by next-generation sequencing (NGS) technologies. Sequencing of reduced representation libraries for eight flax varieties identified about 55 thousand SNPs that could be useful for flax research and breeding (Kumar et al., 2012).

3 GWAS and QTL mapping in flax

Quantitative Trait Locus (QTL) mapping and Genome-Wide Association Study (GWAS) allow the identification of genomic regions responsible for valuable traits. QTL mapping uses biparental populations, while GWAS is based on the analysis of unrelated individuals (Gupta et al., 2019). These approaches can greatly advance the identification of DNA markers for valuable traits and accelerate marker-assisted and genomic selection (Adlak et al., 2023; Veerendrakumar et al., 2024). In 2012, the whole genome of flax was sequenced and assembled (Wang et al., 2012), facilitating GWAS and QTL mapping in this crop. We describe the relevant studies below and summarized the data on the identified QTLs in [Supplementary Table 1](#). [Supplementary Table 1](#) lists QTLs associated with seed-related traits, fiber-related traits, resistance to biotic and abiotic stressors, and some other valuable flax traits. QTL coordinates on the reference genome and references to the works in which the QTLs were identified are also listed.

In many studies, authors identified hundreds or thousands of QTLs associated with different flax traits. In [Supplementary Table 1](#), we included only the most significant QTLs, if they were highlighted by the authors of the studies.

Seed-related traits, which are particularly important for linseed, were in the focus of a significant number of works. QTLs for seed oil content and composition, seed protein content, seed yield and weight, and days to maturity were identified using the recombinant inbred line (RIL) population (Kumar et al., 2015). QTLs for seed oil traits, seed yield, and days to maturity were revealed using genome sequencing of biparental mapping populations (You et al., 2018a). Yield-related QTLs (thousand seed weight, capsule number, number of branches, fatty acid content) were also determined using Specific-Locus Amplified Fragment Sequencing (SLAF-seq) of the flax core collection (Xie et al., 2017, 2018, 2019). In addition, QTLs for seed size and weight were revealed by GWAS of the flax core collection (Guo et al., 2019),

and Quantitative Trait Nucleotides (QTNs) for seed weight were identified using multi-locus GWAS (Sarooha et al., 2023). QTLs associated with fatty acid composition of oil and its content in flax seeds were identified by Genotyping-By-Sequencing (GBS) of the RIL population (Zhao et al., 2020). SNPs associated with seed weight, seed oil content, days to flowering, and plant branching were revealed by GBS of a set of linseed varieties (Singh et al., 2019). Furthermore, QTLs associated with flowering time (Soto-Cerda et al., 2021) and mucilage and hull content (Soto-Cerda et al., 2018) and QTNs associated with flowering and maturity time (Sarooha et al., 2022a) were identified by GWAS analysis of the flax core collections. In addition, SNPs associated with male sterility (Zhao et al., 2023) and QTLs for petal color (Guo et al., 2024) were revealed in flax using GWAS analysis.

The search for QTLs for fiber-related traits is another major direction in flax research, which is particularly important for fiber flax. QTLs for plant height, technical length, and fiber content were identified using SLAF-seq of the flax core collection (Xie et al., 2017, 2018). QTLs for plant height, technical length, straw weight, fiber content, and fiber yield were revealed using genome sequencing of biparental mapping populations (Wu et al., 2018; You et al., 2018a). QTLs for plant height and technical length were also identified using GBS of the RIL population (Zhang et al., 2018). SNPs associated with flax plant type (fiber flax or linseed) were revealed by the analysis of whole-genome sequencing data for the flax core collection (Povkhova et al., 2021). QTNs for fiber traits and plant height were revealed using GWAS (Kanapin et al., 2022; Sarooha et al., 2022a).

In addition to traits related to seed and fiber, the resistance of flax plants to biotic and especially abiotic stressors is of interest. QTNs for drought resistance were identified by GWAS (Soto-Cerda et al., 2020; Sertse et al., 2021). QTLs for salt tolerance were also revealed (Li et al., 2022). QTNs for resistance to Fusarium wilt (*Fusarium oxysporum* f. sp. *lini*) were identified using GWAS (Kanapin et al., 2021) and RIL population analysis (Cloutier et al., 2024). QTLs for resistance to pasmo (*Septoria linicola*) were also revealed using GWAS (He et al., 2019, 2023). In addition, QTLs and QTNs for resistance to powdery mildew (*Oidium lini*) were identified by GWAS (Speck et al., 2022; You et al., 2022).

An important study to systematize the available data on flax QTLs and to determine their location on chromosomes of the variety CDC Bethune was published in 2020 (You and Cloutier, 2020). QTLs associated with flax resistance to biotic and abiotic stressors were summarized in the review of 2022 (Yadav et al., 2022). Fiber-related QTLs were also compiled in the study of 2024 (Gudi et al., 2024). Identification and systematization of flax QTLs are necessary to understand the genetic basis of flax traits and to effectively develop improved varieties using marker-assisted and genomic selection and genome editing. Currently, hundreds of flax genomes were sequenced, and sequencing data are available in databases. However, the lack of available and detailed descriptions of phenotypes of the sequenced samples is now becoming a more pressing problem. Such data would allow comprehensive analyses

to identify the genetic determinants of valuable flax traits and to assess their diversity. In addition, almost all the studies described above used the genome assembly of the variety CDC Bethune (You et al., 2018b) as a reference. However, this genome was assembled without the use of long-read sequencing data and is less complete and accurate than flax genomes obtained later using third-generation sequencing platforms from Oxford Nanopore Technologies or Pacific Biosciences (Sa et al., 2021; Dvorianinova et al., 2022). Therefore, reanalysis of the genome sequencing data for flax samples using a more complete and accurate genome as a reference may improve the results of the QTL search. In addition, the construction of the flax pan-genome could also lead to the efficient localization of QTLs and QTNs. Furthermore, the identification of genes that play a role in the determination of valuable flax traits could be based on transcriptome data. Genes with high expression levels in specific tissues and developmental stages are likely to be involved in the processes occurring there and deserve attention for further research (Galinousky et al., 2020; Dvorianinova et al., 2023; Gorshkova et al., 2023).

4 Functional markers in flax breeding

Many QTLs were identified for valuable flax traits. However, QTLs for the same traits revealed in different studies often do not overlap, suggesting that the studied genotypes, the experimental design, and the data analysis have a significant influence on the obtained results. This hinders the widespread use of QTLs in flax breeding. However, some functional DNA markers, which are not only associated but specifically define valuable flax traits, were identified.

It was shown that nonsense and missense mutations in the *FAD3A* and *FAD3B* genes decrease linolenic acid and increase linoleic acid in linseed oil (Vrinten et al., 2005; Banik et al., 2011; Thambugala et al., 2013; Rajwade et al., 2016; Porokhvinova et al., 2019; Dmitriev et al., 2020). These markers are useful and are already used in linseed breeding to develop food varieties whose oil is more resistant to oxidation compared to traditional linseed varieties (Povkhova et al., 2022).

The researchers also revealed functional SNPs that determine the color of flax seeds. Mutations in the second exon of the *FLAVONOID 3'5' HYDROXYLASE (F3'5'H)* gene negatively affect the synthesis of proanthocyanidins, resulting in a yellow seed coat (Sudarshan et al., 2017). Some of the SNPs in the glutathione S-transferase gene, which lead to four amino acid substitutions, are also likely to result in a yellow color of flax seed coats (Young et al., 2022). These markers are useful for flax breeding because yellow seeds may be more attractive for the use in food products (Abtahi and Mirlohi, 2024).

In addition, *S-lectin receptor-like kinase (SRLK)* (Lus10025891) was proposed as a candidate for a major flax resistance gene to Fusarium wilt. The SNP and indel, which resulted in amino acid substitutions in SRLK, distinguished resistant flax varieties from

susceptible ones and probably determined the resistance (Cloutier et al., 2024). These markers could be used in the development of flax varieties with resistance to the harmful flax pathogen – *Fusarium oxysporum* f. sp. *lini*. Moreover, ethyl methanesulfonate induced a mutation in *LuALS1* that conferred resistance to sulfonylurea herbicides in flax plants (Liu et al., 2023).

5 Prospects of marker-assisted selection in flax

The use of functional DNA polymorphisms in plant breeding requires the utilization of simple and reliable assays. Cleaved Amplified Polymorphic Sequence (CAPS) markers were efficiently used in marker-assisted selection of crops (Shavrukov, 2014, 2016), including flax (Povkhova et al., 2022). However, less labor-intensive and more rapid methods are preferred for the analysis of large sample sets, and approaches based on fluorescence detection are promising. High-Resolution Melting (HRM) allows the identification of DNA polymorphisms based on the comparison of amplicon melting curves and is suitable for high-throughput genotyping of plants (Simko, 2016), including flax (Povkhova et al., 2022). HRM requires a DNA-intercalating dye, as opposed to fluorophore-labeled SNP-specific probes in the TaqMan method (Broccanello et al., 2018; Ayalew et al., 2019). TaqMan assays also allow for rapid and accurate genotyping of large sample sets (Woodward, 2014). Another method for fluorescence detection of DNA polymorphisms is Kompetitive Allele-Specific PCR (KASP) (He et al., 2014), which uses universal Fluorescence Resonance Energy Transfer (FRET) cassettes. FRET cassettes greatly simplify and/or cheapen genotyping, so KASP was used effectively for many crops (Dipta et al., 2024). The KASP markers were involved in QTL mapping and identification of the gene likely responsible for flax resistance to Fusarium wilt (Cloutier et al., 2024). Thus, KASP and similar approaches utilizing FRET cassettes are promising for the identification of SNPs and indels that define valuable flax traits, including in marker-assisted selection.

Microarrays were used extensively in recent years in large-scale studies where samples were needed to be analyzed for a large number of markers simultaneously. The technology allows DNA to be tested for the presence of target sequences by hybridizing to probes attached to a solid platform. This method is convenient for mass use when the same traits are analyzed in many samples, but it is disadvantageous for single use (Yang and Wei, 2015). Studies on flax using microarrays are rare (Roach and Deyholos, 2007; 2008; Fenart et al., 2010).

Targeted amplicon sequencing allows the complete sequence of a gene or locus of interest to be obtained and SNPs/indels to be analyzed (Senapathy et al., 2010; Cronn et al., 2012; Akhmetshina et al., 2020). This approach requires target enrichment, for instance by amplification or hybridization. Targeted sequencing expands horizons in areas such as plant breeding, genetics, evolutionary and phylogenetic studies (Mertes et al., 2011; Cronn et al., 2012; Andermann et al., 2020).

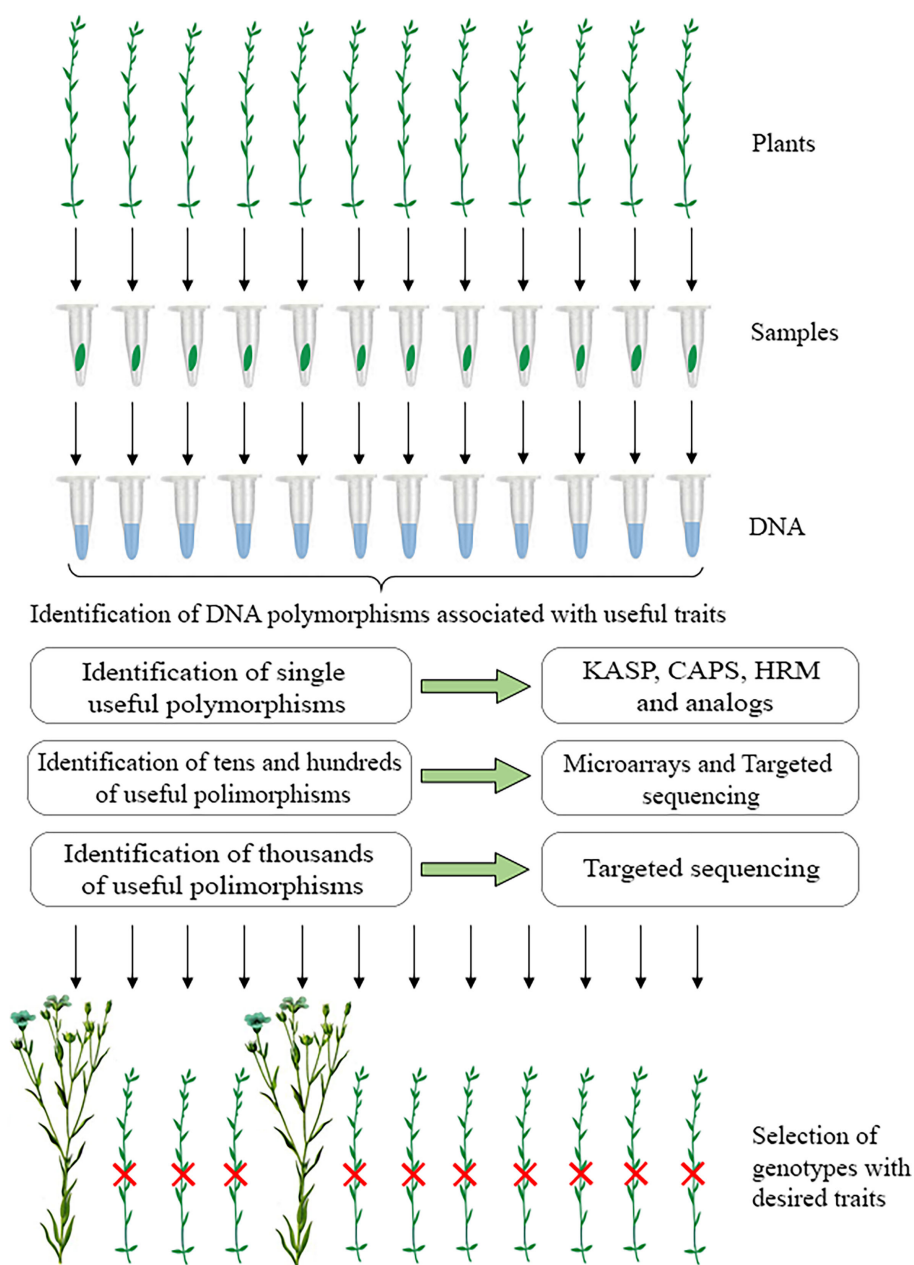


FIGURE 1

A scheme for marker-assisted selection in flax: collection of material from individual plants, isolation of DNA, testing samples for the presence of target DNA polymorphisms using various approaches, and selection of promising plants for breeding based on the obtained results.

The above methods can be applied in flax breeding. A scheme for marker-assisted selection in flax, in which different approaches can be used to identify nucleotides at sites of interest, is shown in Figure 1. It involves collecting material from individual plants, isolating DNA, testing samples for the presence of target DNA polymorphisms, and selecting promising plants for breeding based on the obtained results. The use of these technologies will increase the efficiency of creating improved flax varieties. At the same time, marker-assisted selection in flax requires the identification of a larger number of functional polymorphisms that determine the desired traits. This can be

facilitated by extensive phenotyping of those flax samples for which genome sequencing data are available. As a result, rapid and accurate determination of flax plant traits based on effective DNA test systems will take flax breeding to the next level.

Author contributions

DZ: Writing – original draft, Writing – review & editing. EP: Writing – original draft, Writing – review & editing. TR: Writing –

original draft, Writing – review & editing. EB: Writing – original draft, Writing – review & editing. AA: Writing – original draft, Writing – review & editing. ES: Writing – original draft, Writing – review & editing. ED: Writing – original draft, Writing – review & editing. AD: Writing – original draft, Writing – review & editing. NM: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Part of this work (review of DNA markers of valuable fiber flax traits) was financially supported by the Russian Science Foundation, grant number 24-64-00033, <https://rscf.ru/project/24-64-00033/>. Part of this work (review of DNA markers of valuable linseed traits) was financially supported by the Ministry of Science and Higher Education of the Russian Federation, grant number 075-15-2021-1064.

References

- Abbasi Holasou, H., Abdollahi Mandoulakani, B., Jafari, M., and Bernousi, I. (2016). Use of IRAP and REMAP markers to interpret the population structure of *Linum usitatissimum* from Iran. *Biologia* 71, 305–315. doi: 10.1515/biolog-2016-0042
- Abtahi, M., and Mirlohi, A. (2024). Quality assessment of flax advanced breeding lines varying in seed coat color and their potential use in the food and industrial applications. *BMC Plant Biol.* 24, 60. doi: 10.1186/s12870-024-04733-1
- Adlak, T., Tiwari, S., Rathore, M. S., Tripathi, N., Tiwari, P. N., and Tripathi, M. (2023). Biotechnological approaches for genetic improvement of crops. *book: Cutting Edge Res. Biol.* 7, 63–84. doi: 10.9734/bpi/cebr/v7/5376C
- Adugna, W., Labuschagne, M., and Viljoen, C. (2006). The use of morphological and AFLP markers in diversity analysis of linseed. *Biodiversity Conserv.* 15, 3193–3205. doi: 10.1007/s10531-005-6970-8
- Akhmetshina, A. O., Strygina, K. V., Khlestkina, E. K., Porokhovina, E. A., and Brutch, N. B. (2020). High-throughput sequencing techniques to flax genetics and breeding. *Ecol. Genet.* 18, 103–124. doi: 10.17816/ecogen16126
- Al-Madhagy, S., Ashmawy, N. S., Mamdouh, A., Eldahshan, O. A., and Farag, M. A. (2023). A comprehensive review of the health benefits of flaxseed oil in relation to its chemical composition and comparison with other omega-3-rich oils. *Eur. J. Med. Res.* 28, 240. doi: 10.1186/s40001-023-01203-6
- Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., et al. (2020). A guide to carrying out a phylogenomic target sequence capture project. *Front. Genet.* 10, 1407. doi: 10.3389/fgene.2019.01407
- Asgarinia, P., Cloutier, S., Duguid, S., Rashid, K., Mirlohi, A., Banik, M., et al. (2013). Mapping quantitative trait loci for powdery mildew resistance in flax (*Linum usitatissimum* L.). *Crop Sci.* 53, 2462–2472. doi: 10.2135/cropsci2013.05.0298
- Asyraf, M. R. M., Ishak, M. R., Norrahim, M. N. F., Amir, A. L., Nurazzi, N. M., Ilyas, R. A., et al. (2022). Potential of flax fiber reinforced biopolymer composites for cross-arm application in transmission tower: a review. *Fibers Polymers* 23, 853–877. doi: 10.1007/s12221-022-4383-x
- Ayalew, H., Tsang, P. W., Chu, C., Wang, J., Liu, S., Chen, C., et al. (2019). Comparison of TaqMan, KASP and rhAmp SNP genotyping platforms in hexaploid wheat. *PLoS One* 14, e0217222. doi: 10.1371/journal.pone.0217222
- Banik, M., Duguid, S., and Cloutier, S. (2011). Transcript profiling and gene characterization of three fatty acid desaturase genes in high, moderate, and low linolenic acid genotypes of flax (*Linum usitatissimum* L.) and their role in linolenic acid accumulation. *Genome* 54, 471–483. doi: 10.1139/g11-013
- Bibi, T., Mustafa, H. S. B., Hasan, E., Rauf, S., Mahmood, T., and Ali, Q. (2015). Analysis of genetic diversity in linseed using molecular markers. *Life Sci. J.* 12, 28–37.
- Bo, T., Ma, J., Chen, J., Miao, T., and Zhai, W. (2008). Identification of specific molecular markers linked to the rust resistance gene M4 in flax. *Australas. Plant Pathol.* 37, 417–420. doi: 10.1071/AP08032
- Bo, T. Y., Ye, H. Z., Wang, S. Q., Yang, J. C., Li, X. B., and Zhai, W. X. (2002). Specific molecular markers of the rust resistance gene M4 in flax. *Yi Chuan Xue Bao* 29, 922–927.
- Broccanello, C., Chiodi, C., Funk, A., McGrath, J. M., Panella, L., and Stevanato, P. (2018). Comparison of three PCR-based assays for SNP genotyping in plants. *Plant Methods* 14, 28. doi: 10.1186/s13007-018-0295-6
- Campos, J. R., Severino, P., Ferreira, C. S., Zielinska, A., Santini, A., Souto, S. B., et al. (2019). Linseed essential oil—source of lipids as active ingredients for pharmaceuticals and nutraceuticals. *Curr. Medicinal Chem.* 26, 4537–4558. doi: 10.2174/0929867325666181031105603
- Chandrawati, D., Singh, N., Kumar, R., Kumar, S., Ranade, S., and Kumar Yadav, H. (2017a). Agro-morphological traits and microsatellite marker's based genetic diversity in Indian genotypes of linseed (*Linum usitatissimum* L.). *J. Agricul. Sci. Technol.* 19, 707–718.
- Chandrawati, D., Singh, N., Kumar, R., Kumar, S., Singh, P., Yadav, V., et al. (2017b). Genetic diversity, population structure and association analysis in linseed (*Linum usitatissimum* L.). *Physiol. Mol. Biol. Plants* 23, 207–219. doi: 10.1007/s12298-016-0408-5
- Chandrawati, Maurya, R., Singh, P., Ranade, S., Yadav, H. K. (2014). Diversity analysis in Indian genotypes of linseed (*Linum usitatissimum* L.) using AFLP markers. *Gene* 549, 171–178. doi: 10.1016/j.gene.2014.07.067
- Chen, C., and Liu, Y. (2024). Genetic diversity and distinctness of flax (*Linum usitatissimum* L.) based on morphological and simple sequence repeat (SSR) markers. *Genet. Resour. Crop Evol.* 71, 4763–4777. doi: 10.1007/s10722-024-01933-4
- Choudhary, S. B., Sharma, H. K., Kumar, A. A., Chowdhury, I., Maruthi, R., and Kak, A. (2017a). Genetic diversity spectrum and marker trait association for agronomic traits in global accessions of *Linum usitatissimum* L. *Ind. Crops Products* 108, 604–615. doi: 10.1016/j.indcrop.2017.07.003
- Choudhary, S. B., Sharma, H. K., Kumar, A. A., Maruthi, R. T., Mitra, J., Chowdhury, I., et al. (2017b). SSR and morphological trait based population structure analysis of 130 diverse flax (*Linum usitatissimum* L.) accessions. *Comptes Rendus Biol.* 340, 65–75. doi: 10.1016/j.crvi.2016.12.002
- Cloutier, S., Edwards, T., Zheng, C., Booker, H. M., Islam, T., Nabetani, K., et al. (2024). Fine-mapping of a major locus for Fusarium wilt resistance in flax (*Linum usitatissimum* L.). *Theor. Appl. Genet.* 137, 27. doi: 10.1007/s00122-023-04528-2
- Cloutier, S., Niu, Z., Datla, R., and Duguid, S. (2009). Development and analysis of EST-SSRs for flax (*Linum usitatissimum* L.). *Theor. Appl. Genet.* 119, 53–63. doi: 10.1007/s00122-009-1016-3
- Cloutier, S., Ragupathy, R., Miranda, E., Radovanovic, N., Reimer, E., Walichnowski, A., et al. (2012). Integrated consensus genetic and physical maps of flax (*Linum usitatissimum* L.). *Theor. Appl. Genet.* 125, 1783–1795. doi: 10.1007/s00122-012-1953-0

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.2024.1495069/full#supplementary-material>

- Cloutier, S., Ragupathy, R., Niu, Z., and Duguid, S. (2011). SSR-based linkage map of flax (*Linum usitatissimum* L.) and mapping of QTLs underlying fatty acid composition traits. *Mol. Breed.* 28, 437–451. doi: 10.1007/s11032-010-9494-1
- Corino, C., Rossi, R., Cannata, S., and Ratti, S. (2014). Effect of dietary linseed on the nutritional value and quality of pork and pork products: Systematic review and meta-analysis. *Meat Sci.* 98, 679–688. doi: 10.1016/j.meatsci.2014.06.041
- Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., et al. (2012). Targeted enrichment strategies for next-generation plant biology. *Am. J. Bot.* 99, 291–311. doi: 10.3732/ajb.1100356
- Cullis, C. A., Swami, S., and Song, Y. (1999). RAPD polymorphisms detected among the flax genotypes. *Plant Mol. Biol.* 41, 795–800. doi: 10.1023/a:1006385606163
- Deng, X., Long, S., He, D., Li, X., Wang, Y., Hao, D., et al. (2011). Isolation and characterization of polymorphic microsatellite markers from flax (*Linum usitatissimum* L.). *Afr. J. Biotechnol.* 10, 734–739.
- Diederichsen, A., and Fu, Y.-B. (2006). Phenotypic and molecular (RAPD) differentiation of four infraspecific groups of cultivated flax (*Linum usitatissimum* L. subsp. *usitatissimum*). *Genet. Resour. Crop Evol.* 53, 77–90. doi: 10.1007/s10722-004-0579-8
- Dipta, B., Sood, S., Mangal, V., Bhardwaj, V., Thakur, A. K., Kumar, V., et al. (2024). KASP: a high-throughput genotyping system and its applications in major crop plants for biotic and abiotic stress tolerance. *Mol. Biol. Rep.* 51, 508. doi: 10.1007/s11033-024-09455-z
- Dmitriev, A. A., Kezimana, P., Rozhmina, T. A., Zhuchenko, A. A., Povkhova, L. V., Pushkova, E. N., et al. (2020). Genetic diversity of SAD and FAD genes responsible for the fatty acid composition in flax cultivars and lines. *BMC Plant Biol.* 20, 301. doi: 10.1186/s12870-020-02499-w
- Dvorianinova, E. M., Bolshova, N. L., Pushkova, E. N., Rozhmina, T. A., Zhuchenko, A. A., Novakovskiy, R. O., et al. (2022). Isolating *Linum usitatissimum* L. nuclear DNA enabled assembling high-quality genome. *Int. J. Mol. Sci.* 23, 13244. doi: 10.3390/ijms232113244
- Dvorianinova, E. M., Zinovieva, O. L., Pushkova, E. N., Zhernova, D. A., Rozhmina, T. A., Povkhova, L. V., et al. (2023). Key FAD2, FAD3, and SAD genes involved in the fatty acid synthesis in flax identified based on genomic and transcriptomic data. *Int. J. Mol. Sci.* 24, 14885. doi: 10.3390/ijms241914885
- Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 5, 435–445. doi: 10.1038/nrg1348
- El Sayed, A., Ezzat, S. M., Mostafa, S., Zedan, S., Abdel-Sattar, E., and El Tanbouly, N. (2018). Inter simple sequence repeat analysis of genetic diversity and relationship in four Egyptian flaxseed genotypes. *Pharmacognosy Research.* 10, 166–172. doi: 10.4103/pr.pr_126_17
- Everaert, I., Riek, J. D., Loose, M. D., Waes, J. V., and Bockstaele, E. V. (2001). Most similar variety grouping for distinctness evaluation of flax and linseed (*Linum usitatissimum* L.) varieties by means of AFLP and morphological data. *Plant Varieties and Seeds*, 14, 69–87.
- Fenart, S., Ndong, Y.-P. A., Duarte, J., Rivière, N., Wilmer, J., van Wuytswinkel, O., et al. (2010). Development and validation of a flax (*Linum usitatissimum* L.) gene expression oligo microarray. *BMC Genomics* 11, 592. doi: 10.1186/1471-2164-11-592
- Fombuena, V., Petrucci, R., Dominici, F., Jordá-Vilaplana, A., Montanes, N., and Torre, L. (2019). Maleinized linseed oil as epoxy resin hardener for composites with high bio content obtained from linen byproducts. *Polymers* 11, 301. doi: 10.3390/polym11020301
- Fu, Y.-B. (2005). Geographic patterns of RAPD variation in cultivated flax. *Crop Sci.* 45, 1084–1091. doi: 10.2135/cropsci2004.0345
- Fu, Y.-B., and Allaby, R. G. (2010). Phylogenetic network of *Linum* species as revealed by non-coding chloroplast DNA sequences. *Genet. Resour. Crop Evol.* 57, 667–677. doi: 10.1007/s10722-009-9502-7
- Fu, Y.-B., Diederichsen, A., Richards, K., and Peterson, G. (2002a). Genetic diversity within a range of cultivars and landraces of flax (*Linum usitatissimum* L.) as revealed by RAPDs. *Genet. Resour. Crop Evol.* 49, 167–174. doi: 10.1023/A:1014716031095
- Fu, Y.-B., Guerin, S., Peterson, G. W., Carlson, J. E., and Richards, K. (2003a). Assessment of bulking strategies for RAPD analyses of flax germplasm. *Genet. Resour. Crop Evol.* 50, 743–746. doi: 10.1023/A:1025015401382
- Fu, Y.-B., and Peterson, G. W. (2010). Characterization of expressed sequence tag-derived simple sequence repeat markers for 17 *Linum* species. *Botany* 88, 537–543. doi: 10.1139/B10-019
- Fu, Y.-B., Peterson, G., Diederichsen, A., and Richards, K. (2002b). RAPD analysis of genetic relationships of seven flax species in the genus *Linum* L. *Genet. Resour. Crop Evol.* 49, 253–259. doi: 10.1023/A:1015571700673
- Fu, Y. B., Rowland, G. G., Duguid, S. D., and Richards, K. W. (2003b). RAPD analysis of 54 North American flax cultivars. *Crop Sci.* 43, 1510–1515. doi: 10.2135/cropsci2003.1510
- Galinousky, D., Mokshina, N., Padviski, T., Ageeva, M., Bogdan, V., Kilchevsky, A., et al. (2020). The toolbox for fiber flax breeding: A pipeline from gene expression to fiber quality. *Front. Genet.* 11, 589881. doi: 10.3389/fgene.2020.589881
- Gorshkova, T., Mokshina, N., Mitsuda, N., and Gorshkov, O. (2023). “Key Stages of Flax Bast Fiber Development Through the Prism of Transcriptomics,” in *The Flax Genome*. Eds. F. M. You and B. Fofana (Cham: Springer International Publishing), 149–198.
- Goyal, A., Sharma, V., Upadhyay, N., Gill, S., and Sihag, M. (2014). Flax and flaxseed oil: an ancient medicine & modern functional food. *J. Food Sci. Technol.* 51, 1633–1653. doi: 10.1007/s13197-013-1247-9
- Green, A. G., Chen, Y., Singh, S. P., and Dribnenki, J. C. P. (2008). “Flax”, in *Compendium of Transgenic Crop Plants: Transgenic Oilseed Crops*. Eds. C. Kole and T. C. Hall (Oxford, UK: Blackwell Publishing), 199–226.
- Gudi, S., M, P., Alagappan, P., Raigar, O. P., Halladakeri, P., Gowda, R. S. R., et al. (2024). Fashion meets science: how advanced breeding approaches could revolutionize the textile industry. *Crit. Rev. Biotechnol.* 44, 1653–1679. doi: 10.1080/07388551.2024.2314309
- Guo, D., Jiang, H., and Xie, L. (2024). An R2R3-MYB transcriptional factor luMYB314 associated with the loss of petal pigmentation in flax (*Linum usitatissimum* L.). *Genes (Basel)* 15, 511. doi: 10.3390/genes15040511
- Guo, D., Jiang, H., Yan, W., Yang, L., Ye, J., Wang, Y., et al. (2019). Resequencing 200 flax cultivated accessions identifies candidate genes related to seed size and weight and reveals signatures of artificial selection. *Front. Plant Sci.* 11, 1682. doi: 10.3389/fpls.2019.01682
- Gupta, P. K., Kulwal, P. L., and Jaiswal, V. (2019). Association mapping in plants in the post-GWAS genomics era. *Adv. Genet.* 104, 75–154. doi: 10.1016/b.s.adgen.2018.12.001
- Habibollahi, H., Noormohammadi, Z., Sheidai, M., and Farahani, F. (2015). Genetic structure of cultivated flax (*Linum usitatissimum* L.) based on retrotransposon-based markers. *Genetika* 47, 1111–1122. doi: 10.2298/GENSRI1503111H
- He, C., Holme, J., and Anthony, J. (2014). SNP genotyping: the KASP assay. *Methods Mol. Biol.* 1145, 75–86. doi: 10.1007/978-1-4939-0446-4_7
- He, L., Sui, Y., Che, Y., Wang, H., Rashid, K. Y., Cloutier, S., et al. (2023). Genome-wide association studies using multi-models and multi-SNP datasets provide new insights into pasmo resistance in flax. *Front. Plant Sci.* 14, 1229457. doi: 10.3389/fpls.2023.1229457
- He, L., Xiao, J., Rashid, K. Y., Yao, Z., Li, P., Jia, G., et al. (2019). Genome-wide association studies for pasmo resistance in flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 9, 1982. doi: 10.3389/fpls.2018.01982
- Jhala, A. J., and Hall, L. M. (2010). Flax (*Linum usitatissimum* L.): current uses and future applications. *Aust. J. Basic Appl. Sci.* 4, 4304–4312.
- Kanapin, A., Bankin, M., Rozhmina, T., Samsonova, A., and Samsonova, M. (2021). Genomic regions associated with Fusarium wilt resistance in flax. *Int. J. Mol. Sci.* 22, 12383. doi: 10.3390/ijms222212383
- Kanapin, A., Rozhmina, T., Bankin, M., Surkova, S., Duk, M., Osyagina, E., et al. (2022). Genetic determinants of fiber-associated traits in flax identified by omics data integration. *Int. J. Mol. Sci.* 23, 14536. doi: 10.3390/ijms232314536
- Kezimana, P., Dmitriev, A. A., Kudryavtseva, A. V., Romanova, E. V., and Melnikova, N. V. (2018). Secoisolaricresinol diglucoside of flaxseed and its metabolites: biosynthesis and potential for nutraceuticals. *Front. Genet.* 9, 641. doi: 10.3389/fgene.2018.00641
- Kocak, M. Z., Kaysim, M. G., Aydin, A., Erdinc, C., and Kulak, M. (2023). Genetic diversity of flax genotypes (*Linum usitatissimum* L.) by using agro-morphological properties and molecular markers. *Genet. Resour. Crop Evol.* 70, 2279–2306. doi: 10.1007/s10722-023-01608-6
- Kumar, S., You, F. M., and Cloutier, S. (2012). Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. *BMC Genomics* 13, 684. doi: 10.1186/1471-2164-13-684
- Kumar, S., You, F. M., Duguid, S., Booker, H., Rowland, G., and Cloutier, S. (2015). QTL for fatty acid composition and yield in linseed (*Linum usitatissimum* L.). *Theor. Appl. Genet.* 128, 965–984. doi: 10.1007/s00122-015-2483-3
- Kumari, A., Paul, S., and Sharma, V. (2018). Genetic diversity analysis using RAPD and ISSR markers revealed discrete genetic makeup in relation to fibre and oil content in *Linum usitatissimum* L. genotypes. *Nucleus* 61, 45–53. doi: 10.1007/s13237-017-0206-7
- Lanciková, V., and Žiarovská, J. (2020). Inter-retrotransposon amplified polymorphism markers revealed long terminal repeat retrotransposon insertion polymorphism in flax cultivated on the experimental fields around Chernobyl. *J. Environ. Sci. Health Part A* 55, 957–963. doi: 10.1080/10934529.2020.1760016
- Lawrence, G., Finnegan, J., and Ellis, J. (1993). Instability of the L6 gene for rust resistance in flax is correlated with the presence of a linked Ac element. *Plant J.* 4, 659–669. doi: 10.1046/j.1365-3113x.1993.04040659.x
- Li, X., Guo, D., Xue, M., Li, G., Yan, Q., Jiang, H., et al. (2022). Genome-Wide Association Study of Salt Tolerance at the Seed Germination Stage in Flax (*Linum usitatissimum* L.). *Genes (Basel)* 13, 486. doi: 10.3390/genes13030486
- Liu, C., Zhang, T., Yang, X., Wang, L., Long, Y., Hasi, A., et al. (2023). A luALS mutation with high sulfonylurea herbicide resistance in *Linum usitatissimum* L. *Int. J. Mol. Sci.* 24, 2820. doi: 10.3390/ijms24032820
- Melnikova, N. V., Kudryavtseva, A. V., Zelenin, A. V., Lakunina, V. A., Yurkevich, O. Y., Speranskaya, A. S., et al. (2014). Retrotransposon-based molecular markers for analysis of genetic diversity within the genus *Linum*. *BioMed. Res. Int.* 2014, 231589. doi: 10.1155/2014/231589
- Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M., van der Zaag, P., Franke, A., et al. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. *Briefings Funct. Genomics* 10, 374–386. doi: 10.1093/bfpg/elnr033

- Mhret, W. N., and Heslop-Harrison, J. S. (2018). Biodiversity in Ethiopian linseed (*Linum usitatissimum* L.): molecular characterization of landraces and some wild species. *Genet. Resour. Crop Evol.* 65, 1603–1614. doi: 10.1007/s10722-018-0636-3
- Muravenko, O. V., Lemeshev, V. A., Samatadze, T. E., Amosova, A. V., Grushetskaya, Z. E., Popov, K. V., et al. (2003). Genome comparisons with chromosomal and molecular markers for three closely related flax species and their hybrids. *Russian J. Genet.* 39, 414–421. doi: 10.1023/A:1023309831454
- Muravenko, O. V., Yurkevich, O. Y., Bolsheva, N. L., Samatadze, T. E., Nosova, I. V., Zelenina, D. A., et al. (2009). Comparison of genomes of eight species of sections *Linum* and *Adenolinum* from the genus *Linum* based on chromosome banding, molecular markers and RAPD analysis. *Genetica* 135, 245–255. doi: 10.1007/s10709-008-9273-7
- Nagabhushanam, B., Mir, M. I., Nagaraju, M., Sujatha, E., Devi, B. R., and Kumar, B. K. (2021). Genetic diversity analysis of Linseed (*Linum usitatissimum* L.) accessions using RAPD Markers. *Emirates J. Food Agric. (EJFA)* 33, 589–599. doi: 10.9755/ejfa.2021.v33.i7.2736
- Oh, T., Gorman, M., and Cullis, C. (2000). RFLP and RAPD mapping in flax (*Linum usitatissimum*). *Theor. Appl. Genet.* 101, 590–593. doi: 10.1007/s001220051520
- Osman, M. H., Badawi, M., Zayed, E., Abd El-Maksoud, R. M., and Radwan, K. (2021). Genetic diversity of selected flax genotypes (*Linum usitatissimum* L.) based on scot, issr and ramp markers. *Egyptian J. Genet. And Cytology* 50, 95–117.
- Ottai, M., Al-Kordy, M. A., and Afiah, S. (2011). Evaluation, correlation and path coefficient analysis among seed yield and its attributes of oil flax (*Linum usitatissimum* L.) genotypes. *Australian Journal of Basic and Applied Sciences* 5, 252–258.
- Parikh, M., Maddafor, T. G., Austria, J. A., Aliani, M., Netticadan, T., and Pierce, G. N. (2019). Dietary flaxseed as a strategy for improving human health. *Nutrients* 11, 1171. doi: 10.3390/nu11051171
- Porokhovina, E. A., Shelenga, T. V., Matveeva, T. V., Pavlov, A. V., Grigorieva, E. A., and Brutch, N. B. (2019). Polymorphism of genes controlling low level of linolenic acid in lines from VIR flax genetic collection. *Ecol. Genet.* 17, 5–19. doi: 10.17816/ecogen1725-19
- Povkhova, L. V., Melnikova, N. V., Rozhmina, T. A., Novakovskiy, R. O., Pushkova, E. N., Dvorianinova, E. M., et al. (2021). Genes associated with the flax plant type (oil or fiber) identified based on genome and transcriptome sequencing data. *Plants* 10, 2616. doi: 10.3390/plants10122616
- Povkhova, L. V., Pushkova, E. N., Rozhmina, T. A., Zhuchenko, A. A., Frykin, R. I., Novakovskiy, R. O., et al. (2022). Development and complex application of methods for the identification of mutations in the *FAD3A* and *FAD3B* genes resulting in the reduced content of linolenic acid in flax oil. *Plants (Basel)* 12, 95. doi: 10.3390/plants12010095
- Rachinskaya, O. A., Lemeshev, V. A., Muravenko, O. V., Yurkevich, O. Y., Guzenko, E. V., Bol'sheva, N. L., et al. (2011). Genetic polymorphism of flax *Linum usitatissimum* based on the use of molecular cytogenetic markers. *Russian J. Genet.* 47, 56–65. doi: 10.1134/S1022795411010108
- Ragupathy, R., Rathinavelu, R., and Cloutier, S. (2011). Physical mapping and BAC-end sequence analysis provide initial insights into the flax (*Linum usitatissimum* L.) genome. *BMC Genomics* 12, 217. doi: 10.1186/1471-2164-12-217
- Rajwade, A. V., Arora, R. S., Kadoo, N. Y., Harsulkar, A. M., Ghorpade, P. B., and Gupta, V. S. (2010). Relatedness of Indian flax genotypes (*Linum usitatissimum* L.): an inter-simple sequence repeat (ISSR) primer assay. *Mol. Biotechnol.* 45, 161–170. doi: 10.1007/s12033-010-9256-7
- Rajwade, A. V., Joshi, R. S., Kadoo, N. Y., and Gupta, V. S. (2016). Sequence characterization and in silico structure prediction of fatty acid desaturases in linseed varieties with differential fatty acid composition. *J. Sci. Food Agric.* 96, 4896–4906. doi: 10.1002/jsfa.7775
- Rashid, K., and Duguid, S. (2005). Inheritance of resistance to powdery mildew in flax. *Can. J. Plant Pathol.* 27, 404–409. doi: 10.1080/0706060509507239
- Roach, M. J., and Deyholos, M. K. (2007). Microarray analysis of flax (*Linum usitatissimum* L.) stems identifies transcripts enriched in fibre-bearing phloem tissues. *Mol. Genet. Genomics* 278, 149–165. doi: 10.1007/s00438-007-0241-1
- Roach, M. J., and Deyholos, M. K. (2008). Microarray analysis of developing flax hypocotyls identifies novel transcripts correlated with specific stages of phloem fibre differentiation. *Ann. Bot.* 102, 317–330. doi: 10.1093/aob/mcn110
- Roose-Amsaleg, C., Cariou-Pham, E., Vautrin, D., Tavernier, R., and Solignac, M. (2006). Polymorphic microsatellite loci in *Linum usitatissimum*. *Mol. Ecol. Notes* 6, 796–799. doi: 10.1111/j.1471-8286.2006.01348.x
- Sa, R., Yi, L., Siqin, B., An, M., Bao, H., Song, X., et al. (2021). Chromosome-level genome assembly and annotation of the fiber flax (*Linum usitatissimum*) genome. *Front. Genet.* 12, 735690. doi: 10.3389/fgene.2021.735690
- Saha, D., and Hazra, C. (2004). Agrihistorical genesis of flax (*Linum usitatissimum* LINN.). *J. Agril.* 9, 1–8.
- Saha, D., Rana, R. S., Das, S., Datta, S., Mitra, J., Cloutier, S. J., et al. (2019). Genome-wide regulatory gene-derived SSRs reveal genetic differentiation and population structure in fiber flax genotypes. *J. Appl. Genet.* 60, 13–25. doi: 10.1007/s13353-018-0476-z
- Saini, R. K., Prasad, P., Sreedhar, R. V., Akhilender Naidu, K., Shang, X., and Keum, Y.-S. (2021). Omega-3 polyunsaturated fatty acids (PUFAs): Emerging plant and microbial sources, oxidative stability, bioavailability, and health benefits—A review. *Antioxidants* 10, 1627. doi: 10.3390/antiox10101627
- Saroja, A., Gomashe, S. S., Kaur, V., Pal, D., Ujjainwal, S., Aravind, J., et al. (2023). Genetic dissection of thousand-seed weight in linseed (*Linum usitatissimum* L.) using multi-locus genome-wide association study. *Front. Plant Sci.* 14, 1166728. doi: 10.3389/fpls.2023.1166728
- Saroja, A., Pal, D., Gomashe, S. S., Akash, K., Kaur, V., Ujjainwal, S., et al. (2022a). Identification of QTNs associated with flowering time, maturity, and plant height traits in *Linum usitatissimum* L. using genome-wide association study. *Front. Genet.* 13, 811924. doi: 10.3389/fgene.2022.811924
- Saroja, A., Pal, D., Kaur, V., Kumar, S., Bartwal, A., Aravind, J., et al. (2022b). Agromorphological variability and genetic diversity in linseed (*Linum usitatissimum* L.) germplasm accessions with emphasis on flowering and maturity time. *Genet. Resour. Crop Evol.* 69, 315–333. doi: 10.1007/s10722-021-01231-3
- Schneeberger, R. G., and Cullis, C. A. (1991). Specific DNA alterations associated with the environmental induction of heritable changes in flax. *Genetics* 128, 619–630. doi: 10.1093/genetics/128.3.619
- Senapathy, P., Bhasi, A., Mattox, J., Dhandapany, P. S., and Sadayappan, S. (2010). Targeted genome-wide enrichment of functional regions. *PLoS One* 5, e11138. doi: 10.1371/journal.pone.0011138
- Sertse, D., You, F. M., Ravichandran, S., Soto-Cerda, B. J., Duguid, S., and Cloutier, S. (2021). Loci harboring genes with important role in drought and related abiotic stress responses in flax revealed by multiple GWAS models. *Theor. Appl. Genet.* 134, 191–212. doi: 10.1007/s00122-020-03691-0
- Shavruk, Y. (2014). *Cleaved Amplified Polymorphic Sequences (CAPS) Markers in Plant Biology* (New York, NY, USA: Nova Science Publishers).
- Shavruk, Y. (2016). CAPS markers in plant biology. *Russian J. Genetics: Appl. Res.* 6, 279–287. doi: 10.1134/S2079059716030114
- Simko, I. (2016). High-resolution DNA melting analysis in plant research. *Trends Plant Sci.* 21, 528–537. doi: 10.1016/j.tplants.2016.01.004
- Singh, N., Agarwal, N., and Yadav, H. K. (2019). Genome-wide SNP-based diversity analysis and association mapping in linseed (*Linum usitatissimum* L.). *Euphytica* 215, 1–11. doi: 10.1007/s10681-019-2462-x
- Singh, P., Akram, M., and Srivastava, R. (2009). Genetic diversity in linseed (*Linum usitatissimum*) cultivars based on RAPD-analysis. *Indian J. Agric. Sci.* 79, 1046–1049.
- Singh, N., Kumar, R., Kumar, S., Singh, P. K., and Yadav, H. K. (2021). Mapping QTLs for Alternaria blight in Linseed (*Linum usitatissimum* L.). *3 Biotech.* 11, 91. doi: 10.1007/s13205-020-02638-y
- Singh, K., Mridula, D., Rehal, J., and Barnwal, P. (2011). Flaxseed: a potential source of food, feed and fiber. *Crit. Rev. Food Sci. Nutr.* 51, 210–222. doi: 10.1080/10408390903537241
- Smykal, P., Bačová-Kertesová, N., Kalendar, R., Corander, J., Schulman, A. H., and Pavelek, M. (2011). Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. *Theor. Appl. Genet.* 122, 1385–1397. doi: 10.1007/s00122-011-1539-2
- Soto-Cerda, B. J., Aravena, G., and Cloutier, S. (2021). Genetic dissection of flowering time in flax (*Linum usitatissimum* L.) through single- and multi-locus genome-wide association studies. *Mol. Genet. Genomics* 296, 877–891. doi: 10.1007/s00438-021-01785-y
- Soto-Cerda, B. J., Cloutier, S., Gajardo, H. A., Aravena, G., and Quian, R. (2019). Identifying drought-resilient flax genotypes and related-candidate genes based on stress indices, root traits and selective sweep. *Euphytica* 215, 1–16. doi: 10.1007/s10681-019-2362-0
- Soto-Cerda, B. J., Cloutier, S., Gajardo, H. A., Aravena, G., Quian, R., and You, F. M. (2020). Drought response of flax accessions and identification of quantitative trait nucleotides (QTNs) governing agronomic and root traits by genome-wide association analysis. *Mol. Breed.* 40, 1–24. doi: 10.1007/s11032-019-1096-y
- Soto-Cerda, B. J., Cloutier, S., Quian, R., Gajardo, H. A., Olivos, M., and You, F. M. (2018). Genome-wide association analysis of mucilage and hull content in flax (*Linum usitatissimum* L.) seeds. *Int. J. Mol. Sci.* 19, 2870. doi: 10.3390/ijms19102870
- Soto-Cerda, B. J., Diederichsen, A., Duguid, S., Booker, H., Rowland, G., and Cloutier, S. (2014). The potential of pale flax as a source of useful genetic variation for cultivated flax revealed through molecular diversity and association analyses. *Mol. Breed.* 34, 2091–2107. doi: 10.1007/s11032-014-0165-5
- Soto-Cerda, B. J., Diederichsen, A., Ragupathy, R., and Cloutier, S. (2013). Genetic characterization of a core collection of flax (*Linum usitatissimum* L.) suitable for association mapping studies and evidence of divergent selection between fiber and linseed types. *BMC Plant Biol.* 13, 78. doi: 10.1186/1471-2229-13-78
- Soto-Cerda, B. J., Maureira-Butler, I., Muñoz, G., Rupayan, A., and Cloutier, S. (2012). SSR-based population structure, molecular diversity and linkage disequilibrium analysis of a collection of flax (*Linum usitatissimum* L.) varying for mucilage seed-coat content. *Mol. Breed.* 30, 875–888. doi: 10.1007/s11032-011-9670-y
- Soto-Cerda, B. J., Saavedra, H. U., Navarro, C. N., and Ortega, P. M. (2011). Characterization of novel genic SSR markers in *Linum usitatissimum* (L.) and their transferability across eleven *Linum* species. *Electronic J. Biotechnol.* 14, 4–4. doi: 10.2225/vol14-issue2-fulltext-6
- Speck, A., Trouve, J. P., Enjalbert, J., Geffroy, V., Joets, J., and Moreau, L. (2022). Genetic architecture of powdery mildew resistance revealed by a genome-wide association study of a worldwide collection of flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 13, 871633. doi: 10.3389/fpls.2022.871633

- Spielmeier, W., Green, A., Bittisnich, D., Mendham, N., and Lagudah, E. (1998). Identification of quantitative trait loci contributing to Fusarium wilt resistance on an AFLP linkage map of flax (*Linum usitatissimum*). *Theor. Appl. Genet.* 97, 633–641. doi: 10.1007/s001220050939
- Stegni, V., IuV, C., and Salina, E. (2000). RAPD analysis of the flax (*Linum usitatissimum* L.) varieties and hybrids of various productivity. *Genetika* 36, 1370–1373.
- Sudarshan, G. P., Kulkarni, M., Akhova, L., Ashe, P., Shaterian, H., Cloutier, S., et al. (2017). QTL mapping and molecular characterization of the classical *D* locus controlling seed and flower color in *Linum usitatissimum* (flax). *Sci. Rep.* 7, 15751. doi: 10.1038/s41598-017-11565-7
- Thambugala, D., Duguid, S., Loewen, E., Rowland, G., Booker, H., You, F. M., et al. (2013). Genetic variation of six desaturase genes in flax and their impact on fatty acid composition. *Theor. Appl. Genet.* 126, 2627–2641. doi: 10.1007/s00122-013-2161-2
- Uysal, H., Fu, Y.-B., Kurt, O., Peterson, G. W., Diederichsen, A., and Kusters, P. (2010). Genetic diversity of cultivated flax (*Linum usitatissimum* L.) and its wild progenitor pale flax (*Linum bienne* Mill.) as revealed by ISSR markers. *Genet. Resour. Crop Evol.* 57, 1109–1119. doi: 10.1007/s10722-010-9551-y
- Van Treuren, R., Van Soest, L., and Van Hintum, T. J. (2001). Marker-assisted rationalisation of genetic resource collections: a case study in flax using AFLPs. *Theor. Appl. Genet.* 103, 144–152. doi: 10.1007/s001220100537
- Veerendrakumar, H., Barmukh, R., Shah, P., Bomireddy, D., Jamedar, H. V. R., Roorkiwal, M., et al. (2024). “Genomic Selection in Crop Improvement,” in *Frontier Technologies for Crop Improvement* Eds. M. K. Pandey, A. Bentley, H. Desmae, M. Roorkiwal and R. K. Varshney (Singapore: Springer, Singapore), 187–222.
- Vrinten, P., Hu, Z., Munchinsky, M.-A., Rowland, G., and Qiu, X. (2005). Two *FAD3* desaturase genes control the level of linolenic acid in flax seed. *Plant Physiol.* 139, 79–87. doi: 10.1104/pp.105.064451
- Wakjira, A., Viljoen, C., and Labuschagne, M. (2005). Analysis of genetic diversity in linseed using AFLP markers. *SINET: Ethiopian J. Sci.* 28, 41–50. doi: 10.4314/sinet.v28i1.18230
- Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., et al. (2012). The genome of flax (*Linum usitatissimum*) assembled *de novo* from short shotgun sequence reads. *Plant J.* 72, 461–473. doi: 10.1111/j.1365-3113X.2012.05093.x
- Wiesnerova, D., and Wiesner, I. (2004). ISSR-based clustering of cultivated flax germplasm is statistically correlated to thousand seed mass. *Mol. Biotechnol.* 26, 207–214. doi: 10.1385/MB:26:3:207
- Woodward, J. (2014). Bi-allelic SNP genotyping using the TaqMan® assay. *Crop Breeding: Methods Protoc.* Eds. D. Fleury and R. Whitford (New York, NY, USA: Springer), 67–74. doi: 10.1007/978-1-4939-0446-4_6
- Wu, J., Zhao, Q., Wu, G., Zhang, S., and Jiang, T. (2017). Development of novel SSR markers for flax (*Linum usitatissimum* L.) using reduced-representation genome sequencing. *Front. Plant Sci.* 7, 2018. doi: 10.3389/fpls.2016.02018
- Wu, J., Zhao, Q., Zhang, L., Li, S., Ma, Y., Pan, L., et al. (2018). QTL mapping of fiber-related traits based on a high-density genetic map in flax (*Linum usitatissimum* L.). *Front. Plant Sci.* 9, 885. doi: 10.3389/fpls.2018.00885
- Xie, D., Dai, Z., Yang, Z., Sun, J., Zhao, D., Yang, X., et al. (2017). Genome-wide association study identifying candidate genes influencing important agronomic traits of flax (*Linum usitatissimum* L.) using SLAF-seq. *Front. Plant Sci.* 8, 2232. doi: 10.3389/fpls.2017.02232
- Xie, D., Dai, Z., Yang, Z., Tang, Q., Deng, C., Xu, Y., et al. (2019). Combined genome-wide association analysis and transcriptome sequencing to identify candidate genes for flax seed fatty acid metabolism. *Plant Sci.* 286, 98–107. doi: 10.1016/j.plantsci.2019.06.004
- Xie, D., Dai, Z., Yang, Z., Tang, Q., Sun, J., Yang, X., et al. (2018). Genomic variations and association study of agronomic traits in flax. *BMC Genomics* 19, 512. doi: 10.1186/s12864-018-4899-z
- Yadav, B., Kaur, V., Narayan, O. P., Yadav, S. K., Kumar, A., and Wankhede, D. P. (2022). Integrated omics approaches for flax improvement under abiotic and biotic stress: Current status and future prospects. *Front. Plant Sci.* 13, 931275. doi: 10.3389/fpls.2022.931275
- Yang, C., and Wei, H. (2015). Designing microarray and RNA-Seq experiments for greater systems biology discovery in modern plant genomics. *Mol. Plant* 8, 196–206. doi: 10.1016/j.molp.2014.11.012
- You, F. M., and Cloutier, S. (2020). Mapping quantitative trait loci onto chromosome-scale pseudomolecules in flax. *Methods Protoc.* 3, 28. doi: 10.3390/mps3020028
- You, F. M., Moumen, I., Khan, N., and Cloutier, S. (2023). “Reference genome sequence of flax,” in *The Flax Genome*. Eds F. M. You and B. Fofana (Cham, Switzerland: Springer), 1–17.
- You, F. M., Rashid, K. Y., Zheng, C., Khan, N., Li, P., Xiao, J., et al. (2022). Insights into the Genetic Architecture and Genomic Prediction of Powdery Mildew Resistance in Flax (*Linum usitatissimum* L.). *Int. J. Mol. Sci.* 23, 4960. doi: 10.3390/ijms23094960
- You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018a). Genome-wide association study and selection signatures detect genomic regions associated with seed yield and oil quality in flax. *Int. J. Mol. Sci.* 19, 2303. doi: 10.3390/ijms19082303
- You, F. M., Xiao, J., Li, P., Yao, Z., Jia, G., He, L., et al. (2018b). Chromosome-scale pseudomolecules refined by optical, physical and genetic maps in flax. *Plant J.* 95, 371–384. doi: 10.1111/tpj.2018.95.issue-2
- Young, L., Akhova, L., Kulkarni, M., You, F., and Booker, H. (2022). Fine-mapping of a putative glutathione S-transferase (GST) gene responsible for yellow seed colour in flax (*Linum usitatissimum*). *BMC Res. Notes* 15, 72. doi: 10.1186/s13104-022-05964-x
- Zhang, J., Long, Y., Wang, L., Dang, Z., Zhang, T., Song, X., et al. (2018). Consensus genetic linkage map construction and QTL mapping for plant height-related traits in linseed flax (*Linum usitatissimum* L.). *BMC Plant Biol.* 18, 160. doi: 10.1186/s12870-018-1366-6
- Zhao, X., Yi, L., Zuo, Y., Gao, F., Cheng, Y., Zhang, H., et al. (2023). High-quality genome assembly and genome-wide association study of male sterility provide resources for flax improvement. *Plants (Basel)* 12, 2773. doi: 10.3390/plants12152773
- Zhao, L., Zhao, W., Wang, L., Wang, B., Dang, Z., and Li, W. (2020). Construction of A High-Density SNP Genetic Map by Genotyping-By-Sequencing and Location of Seed Quality traits in Flax (*Linum usitatissimum* L.). *Res. Square*. doi: 10.21203/rs.3.rs-70683/v1
- Žiarovská, J., Špeváková, I., Klongová, L., Farkasová, S., and Rashydow, N. (2022). Transposable elements in the revealing of polymorphism-based differences in the seeds of flax varieties grown in remediated chernobyl. *Plants* 11, 2567. doi: 10.3390/plants11192567



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Gograj Singh Jat,
Division of Vegetable Science - IARI, India
Akanksha Singh,
Purdue University, United States

*CORRESPONDENCE

Aurora Díaz

✉ adiazb@cita-aragon.es;

✉ adiazb@unizar.es

RECEIVED 10 September 2024

ACCEPTED 23 December 2024

PUBLISHED 15 January 2025

CITATION

Medina-Lozano I, Grimplet J and Díaz A
(2025) Harnessing the diversity of a
lettuce wild relative to identify anthocyanin-
related genes transcriptionally responsive to
drought stress.

Front. Plant Sci. 15:1494339.

doi: 10.3389/fpls.2024.1494339

COPYRIGHT

© 2025 Medina-Lozano, Grimplet and Díaz.

This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Harnessing the diversity of a lettuce wild relative to identify anthocyanin-related genes transcriptionally responsive to drought stress

Inés Medina-Lozano^{1,2}, Jérôme Grimplet^{1,2} and Aurora Díaz^{1,2*}

¹Department of Plant Sciences, Agrifood Research and Technology Centre of Aragon (CITA), Zaragoza, Spain, ²AgriFood Institute of Aragon – IA2 (CITA-University of Zaragoza), Zaragoza, Spain

Lettuce is a crop particularly vulnerable to drought. A transcriptomic study in the variety 'Romired' and the wild relative *Lactuca homblei* was conducted to understand the increase in anthocyanins (only significant in *L. homblei*) in response to drought previously observed. RNA-seq revealed more differentially expressed genes (DEGs), especially upregulated, in the wild species, in which the most abundant and significant GO terms were involved in regulatory processes (including response to water). Anthocyanin synthesis was triggered in *L. homblei* in response to drought, with 17 genes activated out of the 36 mapped in the phenylpropanoid-flavonoid pathway compared to 7 in 'Romired'. Nineteen candidate DEGs with the strongest change in expression and correlation with both anthocyanin content and drought were selected and validated by qPCR, all being differentially expressed only in the wild species with the two techniques. Their functions were related to anthocyanins and/or stress response and they harboured 404 and 11 polymorphisms in the wild and cultivated species, respectively. Some wild variants had high or moderate predicted impacts on the respective protein function: a transcription factor that responds to abiotic stresses, a heat shock protein involved in stomatal closure, and a phospholipase participating in anthocyanin accumulation under abiotic stress. These genetic variants could explain the differences in the gene expression patterns between the wild (significantly up/downregulated) and the cultivated (no significant changes) species. The diversity of this crop wild relative for anthocyanin-related genes involved in the response to drought could be exploited to improve lettuce resilience against some adverse climate effects.

KEYWORDS

abiotic stress, antioxidants, crop wild relatives, differentially expressed genes, *Lactuca sativa* L., real-time qPCR, RNA-seq, resilience

1 Introduction

Abiotic stress is a major challenge for agriculture, especially in the present scenario of climate change (IPCC, 2021), in which adverse environmental conditions are more and more frequent (FAO, 2021). In particular, drought is one of the most concerning abiotic stresses, affecting both crop yield and quality. Drought stress has important effects on plant growth by affecting diverse physiological and biochemical processes, like cell expansion and photosynthesis due to stomatal closure (Farooq et al., 2009). Apart from biomass production, it also affects nutrient composition and concentration as well as secondary metabolism, depending generally on the stress severity and duration, as well as on plant tolerance (Reddy et al., 2004; Medina-Lozano et al., 2024).

Lettuce (*Lactuca sativa* L.) is one of the most important leafy vegetables worldwide (FAOSTAT, 2021). It provides different health benefits attributed to phenolic compound, vitamin, and fibre contents (Llorach et al., 2008), among others, what contributes to increase its popularity especially with the growing awareness of the impact of diet on health among consumers. Lettuce is mostly composed by water (up to 97%) (Mou, 2005), what makes it highly susceptible to drought (Eriksen et al., 2016). However, controlled deficit irrigation can cause an improvement of its health-promoting properties by increasing the content of some antioxidants (Paim et al., 2020; Medina-Lozano et al., 2024). Among the phenolic compounds present in lettuce, anthocyanins are responsible for red pigmentation of the leaves in semi-red and red varieties. They are known to play crucial roles in human health due to their antioxidant properties (Garcia and Blesso, 2021). It has been described that water stress causes an accumulation of anthocyanins in some fruits, vegetables and oil crops, such as grapes (Ju et al., 2019), strawberries (Rugienius et al., 2021), and purple-stem canola (Chen et al., 2022b). In lettuce, different studies had reported increased levels of either total phenolic compounds under water stress (Zeljkočić et al., 2023), or anthocyanins in response to other environmental stresses, like UV irradiance (Tsormpatsidis et al., 2008) and low temperatures (Becker et al., 2014). However, anthocyanin response to drought conditions had barely been studied in this crop until recently, when a drought-induced anthocyanin accumulation not only in cultivated lettuce

varieties but also in wild relative species, has been discovered (Medina-Lozano et al., 2024).

Lettuce anthocyanin content is very dependent on the genotype. In absence of stress, commercial varieties are the richest, followed by traditional ones and finally by lettuce wild relatives (Medina-Lozano et al., 2021). Interestingly, in all the lettuce-related germplasm studied, the water stress always resulted in an increase of the total anthocyanin content, with the highest accumulation detected in a wild relative species (Medina-Lozano et al., 2024). Crop wild relatives (CWR) are known to be a source of favourable alleles for interesting traits for breeding, like resistance to diseases or tolerance to abiotic stresses (Quezada-Martinez et al., 2021).

Unveiling the molecular mechanisms governing the changes of anthocyanin content in response to water stress could have multiple benefits from a breeding perspective, aiming at enhancing the drought tolerance of the crop and the antioxidant properties of the food product. In lettuce, the great majority of transcriptomic studies related to anthocyanins are focused on the differences between green and red varieties (Moreno-Escamilla et al., 2020; Su et al., 2020). RNA-seq has also been used to study different abiotic stresses in this crop, e.g., high and low temperatures (Park et al., 2020; Chen et al., 2022a), the presence of heavy metals (Xiong et al., 2021), and even drought (Koyama et al., 2021). However, the specific effect of environmental factors on lettuce anthocyanin regulation has been scarcely studied, except in the case of different light conditions (Zhang et al., 2018; Wada et al., 2022).

Nowadays, RNA-seq is the most widely used technology for studying gene expression due to its many advantages. RNA-seq is a precise and sensitive technique that has also a wide range of detection and is highly accurate in terms of quantification (Wang et al., 2009). Despite being a powerful technique, some artefacts may be present in RNA-seq data (Everaert et al., 2017). Therefore, their validation with an independent technique like real-time quantitative PCR (qPCR) is advisable and even necessary when genes are small, have few exons or low levels of expression (Everaert et al., 2017).

Once differentially expressed genes (DEGs) have been identified, the study of polymorphisms in their sequences might provide information about functional and structural effects that could explain the observed variation for the trait of interest. However, the elucidation of these effects through experimental approaches is usually time and labour consuming and, in many cases, leads to dead ends. That is why the development and use of computational prediction tools as a first approach have experienced a boom in the last few years as they are able to provide increasingly more accurate information to assess phenotypic effects (Yazar and Özbek, 2021).

Metabolite-mediated drought adaptation is an emerging subject that has revealed the importance of some primary metabolites, such as sugars, small peptides, and amino acids, among others, in plant response, either acting as signal factors or as protectors (Zhang et al., 2024). Less is known about the participation of secondary metabolites (e.g., anthocyanins) in plant response to drought, beyond their antioxidant activity like scavengers of reactive oxygen species (ROS) (Naing and Kim, 2021). In this work, we have carried out transcriptomic analyses via RNA-seq and real-time qPCR in a red lettuce variety and a wild relative species that

Abbreviations: ABA, abscisic acid; C, control; cDNA, complementary deoxyribonucleic acid; CHS, chalcone synthase; CITA, Agrifood Research and Technology Centre of Aragón; CPM, counts per million; CWR, crop wild relatives; DEG, differentially expressed gene; DI, deficit irrigation; EBG, early biosynthesis gene; FC, fold change; FDR, false discovery rate; GO, gene ontology; HSP, heat shock protein; Indel, insertion-deletion; LBG, late biosynthesis gene; MNP, multiple nucleotide polymorphisms; MYB, myeloblastosis; NAC, NAM (no apical meristem), ATAF (*Arabidopsis thaliana* activating factor), and CUC (cup-shaped cotyledon); PLIP, phospholipid-inositol phosphatase; PTFE, polytetrafluoroethylene; qPCR, quantitative polymerase chain reaction; RNA-seq, ribonucleic acid sequencing; ROS, reactive oxygen species; RTL3, ribonuclease III-like protein 3; siRNA, small interfering RNA; SNP, single nucleotide polymorphism; spp., species; TF, transcription factor; Vs., versus; WGCNA, weighted gene co-expression network analysis.

experienced a raise in anthocyanin content as a response to drought stress (Medina-Lozano et al., 2024). In addition, *in silico* predictions of the effects of polymorphisms in DEGs could potentially explain the observed differences between the two species in anthocyanin content in plants subject to water stress. The genetic knowledge of this response is key to obtaining new lettuce varieties with both enhanced drought tolerance and health-promoting properties, at the same time that water resources destined to irrigation could be cut down.

2 Materials and methods

2.1 Plant material

Two different accessions of the genus *Lactuca* were included in this study: a commercial variety, the red-leaf lettuce ‘Romired’, and a wild relative species, *Lactuca homblei* De Wild. They were selected from a previous drought stress experiment in which two irrigation regimes, control (C, week 1: 1350 mL, weeks 2-3: 2100 mL/each) and deficit irrigation (DI, weeks 1-3: 0 mL), were tested in two consecutive years (Medina-Lozano et al., 2024). Three biological replicates for the two accessions in each of the two conditions (C and DI) from the experiment carried out in winter 2020/2021 were used to proceed with the transcriptomic studies.

2.2 RNA extraction and sequencing, data processing and DEG identification

Total RNA extraction from lyophilized samples coming from 12 samples (2 accessions x 2 irrigation regimes x 3 biological replicates) was performed using the NZY Total RNA Isolation kit (NZYtech Lda.-Genes and Enzymes, Lisbon, Portugal) as described before (Medina-Lozano et al., 2023). RNA was treated with DNase using the TURBO DNA-free™ kit (Invitrogen, Waltham, MA, USA), following the manufacturers’ instructions. RNA quantity and purity were assessed with the Eukaryotic Total RNA Nanobioanalyzer Assay in a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

The obtained RNA samples from ‘Romired’ and *L. homblei* were used to perform the RNA-seq. They were processed to build a total of 12 strand-specific cDNA libraries. Sequencing of the libraries was performed in both directions with a NovaSeq 6000 S1 instrument (Illumina, San Diego, CA, USA) using the TruSeq Stranded mRNA protocol (Illumina) to obtain between 36 and 111 strand-specific pair-end reads of 100 base pair (bp) lengths per sample. Sequencing was carried out at the National Centre for Genomic Regulation (CNAG-CRG, Barcelona, Spain).

Sequences were analysed using the Galaxy tool (The Galaxy Community, 2022). Adapter sequences were removed by processing the reads from the 12 individual datasets using Trimmomatic (Galaxy version 0.38.1) (Bolger et al., 2014). RNA-seq data alignment to the lettuce reference genome *Lactuca sativa* ‘Salinas’ v8 (Reyes-Chin-Wo et al., 2017) was performed using HISAT2

(Galaxy version 2.2.1+galaxy0) (Kim et al., 2015), with a maximum intron length set at 20,000 bp. The Picard tools (<http://broadinstitute.github.io/picard>) MarkDuplicates (Galaxy version 2.18.2.2) and FixMateInformation (Galaxy version 2.18.2.1) were used to filter out the optical duplicates and to mate-pairs, respectively. featureCounts (Galaxy version 2.0.1+galaxy2) (Liao et al., 2013) was used to generate read counts using the gene annotation available in the literature (Reyes-Chin-Wo et al., 2017).

Analysis of differential gene expression between treatments (C and DI) within each of the two accessions was conducted using edgeR (Galaxy version 3.36.0+galaxy0) (Robinson et al., 2009). Genes were considered to be differentially expressed when values of $|\log_2(\text{FC, fold change})| > 1$ and FDR (False Discovery Rate) < 0.05 (adjusted *p*-value via the Benjamini-Hochberg method).

2.3 Structural and functional analysis of the DEGs

Venn diagrams were performed with DEG datasets using the R stats package VennDiagram (<https://CRAN.R-project.org/package=VennDiagram>). GO (Gene Ontology) enrichment analyses were carried out using the tool GOEnrichment from Galaxy platform (Galaxy version 2.0.1) (Faria, 2017), with *p*-value cut-off < 0.05 and using Benjamini-Hochberg multiple test correction. Enriched GO terms involving three categories, biological processes, cellular components, and molecular functions, were evaluated. GO terms of the DEGs were obtained from predicted data using information available in the literature (Reyes-Chin-Wo et al., 2017). Heatmaps were constructed using gplots (<https://CRAN.R-project.org/package=gplots>) and ggplot2 (Wickham, 2009) R stats packages.

2.4 Selection of DEGs

The selection of genes for expression data validation was based on different criteria. First, genes were filtered out for values of $|\log_2(\text{FC})| > 4$ and $\text{FDR} < 0.05$ (substantial increase or decrease in expression levels). Among them, those exhibiting high and significant correlation (both positive and negative) with, first, anthocyanin content and, second, drought stress treatment, were selected. Finally, gene functions related to both anthocyanin content and/or response to different stresses were also taken into account for the selection of a total of 19 DEGs.

Correlations between gene expression and both anthocyanin content and treatments were established through weighted gene co-expression network analysis (WGCNA), which was conducted using the R stats package WGCNA (Langfelder and Horvath, 2008). Normalised RNA-seq data of all genes were used for the WGCNA, except for those with a very low expression among the no DEGs (i.e., less than 5 reads per sample in the three biological samples of each group (C and DI)). Data from both species, *L. sativa* (cultivated lettuce ‘Romired’) and *L. homblei*, were analysed separately.

2.5 DEG validation using real-time qPCR

Total RNA was extracted from each of the 12 samples described above. Subsequently, mRNA was purified using the Dynabeads mRNA DIRECT™ kit (Invitrogen) and cDNA was synthesized using the NZY M-MuLV First-Strand cDNA Synthesis, separate oligos kit (NZYTech) as described before (Medina-Lozano et al., 2023).

Specific pairs of primers for each of the 19 selected DEGs (Supplementary Table S1) were designed using OLIGO software version 6.45 (Cascade, CO, USA) from a consensus sequence of the two species under study, *L. sativa* and *L. homblei*, excluding any ambiguity in the sequences. Real-time qPCR reactions were performed on a StepOnePlus™ System (Applied Biosystems, Waltham, MA, USA) with two technical replicates per each of the three biological replicates. Each reaction contained 1 µL of 1:5 diluted cDNA, 0.40 µM of forward and reverse primers (Integrated DNA Technologies, IDT, Coralville, Iowa, USA), and 1x NZYSupreme qPCR Green Master Mix, ROX plus (NZYTech) in a final volume of 11 µL. The amplification conditions were: 2 min at 95°C, 40 cycles of 5 s at 95°C, 15 s at 52–66°C (Supplementary Table S1) and 30 s at 72°C, followed by the melting curve analysis that ranged from 72°C to 95°C with 0.3°C increment per cycle to verify that a single product was amplified. Non-template controls were included to ensure that contamination with genomic DNA had not occurred.

TRXL3-3 was used as reference gene to normalise qPCR data (Medina-Lozano et al., 2023). Relative expression levels were obtained using the Pfaffl method (Pfaffl, 2001) with some modifications: arithmetic instead of geometric mean was calculated due to the presence of zero values in the raw data (either genes completely shut down as a consequence of the DI or the other way round, unexpressed genes in C conditions that were activated with the DI). This explains values different from 1 in C samples and why they have been represented separately from the DI data in qPCR results.

Student *t*-test was used to assess whether the differences between the means from the qPCR expression data of samples under C and DI conditions were statistically significant. Data transformations ($1/(1+x)^2$ or $1/\sqrt{(x+1)}$) were applied when needed to achieve a normal distribution. Alternatively, Wilcoxon test was used with non-normally distributed data. Statistical analyses were conducted using the software JMP v5.1.2 for Windows (SAS Institute Inc. Cary, NC).

2.6 Polymorphism search, annotation, and effect prediction in the DEGs

Detection of polymorphisms was carried out using the sequences of the 12 samples aligned to the lettuce reference genome (Reyes-Chin-Wo et al., 2017) and processed as explained in subsection 2.2. Firstly, variant calling was performed using FreeBayes package (Galaxy Version 1.3.6+galaxy0) (Tange, 2011; Garrison and Marth, 2012) from Galaxy platform. Then, VCFfilter (Galaxy Version 1.0.0_rc3+galaxy3) (Garrison, 2015) was used to

remove polymorphisms with a total read depth at the locus < 10, QUAL < 20, and number of alternative alleles in called genotypes > 0. In addition, those polymorphic sites exhibiting different genotypes among the total number of samples within accessions and/or more than two different genotypes in comparison with the reference genome in more than 70% of the cases, were filtered out with Excel. Any possible ambiguous polymorphism was also eliminated.

The effect of each polymorphism was annotated and predicted using the SnpEff tool (Galaxy Version 4.3+T.galaxy1) and a snpEff database created using the SnpEff build tool (Galaxy Version 4.3+T.galaxy4) (Cingolani et al., 2012) from the annotation dataset and the FASTA file of *L. sativa* ‘Salinas’ v8 (Reyes-Chin-Wo et al., 2017).

3 Results

3.1 Transcriptome analysis

To investigate the involvement of anthocyanins at molecular level in the response mechanism to drought stress of *Lactuca* spp., a transcriptomic analysis via RNA-seq was performed using plants belonging to the CWR *L. homblei* and to the red commercial lettuce variety ‘Romired’ coming from a previous experiment carried out in winter 2020/2021 (Medina-Lozano et al., 2024). Samples of both accessions showed an accumulation of anthocyanins under DI in comparison to C conditions, though the differences only resulted statistically significant in the case of the wild species *L. homblei* (Medina-Lozano et al., 2024). In particular, three different anthocyanins were identified: cyanidin 3-O-(6'-O-malonylglucoside) was the predominant one and was detected in both accessions and treatments; peonidin 3-O-glucoside appeared under both treatments in the commercial variety, but only under DI in the CWR; and cyanidin 3-(6''-acetylglucoside), exclusively identified under DI conditions in the commercial variety.

After processing the data from *L. homblei*, the clean reads ranged from 40.76 to 51.75 Gb and the percentage of uniquely mapped sequences to the reference genome ranged from 32.39% to 37.40%. In the case of ‘Romired’, the clean reads ranged from 35.43 to 110.62 Gb and the uniquely mapped sequences from 81.04% to 84.81% (Table 1). RNA-seq data from both accessions were aligned to the lettuce reference genome *L. sativa* ‘Salinas’ v8 (Reyes-Chin-Wo et al., 2017). However, *L. homblei* belongs to the tertiary lettuce gene pool (PGR (Plant Genetic Resources) Lettuce, <https://www.pgrportal.nl/en/lettuce-geneticresources-portal.htm>), so it is quite distant from *L. sativa*, what might explain its lower values in terms of uniquely mapped sequences.

3.2 Identification and analysis of DEGs under drought stress conditions

A total of 6,179 DEGs were identified when *L. homblei* plants under C and DI treatments were compared (3,113 upregulated and 3,066 downregulated genes), whereas a total of 5,329 DEGs were obtained in ‘Romired’ plants for the same treatments (1,747

TABLE 1 Statistical summary of RNA-sequencing data.

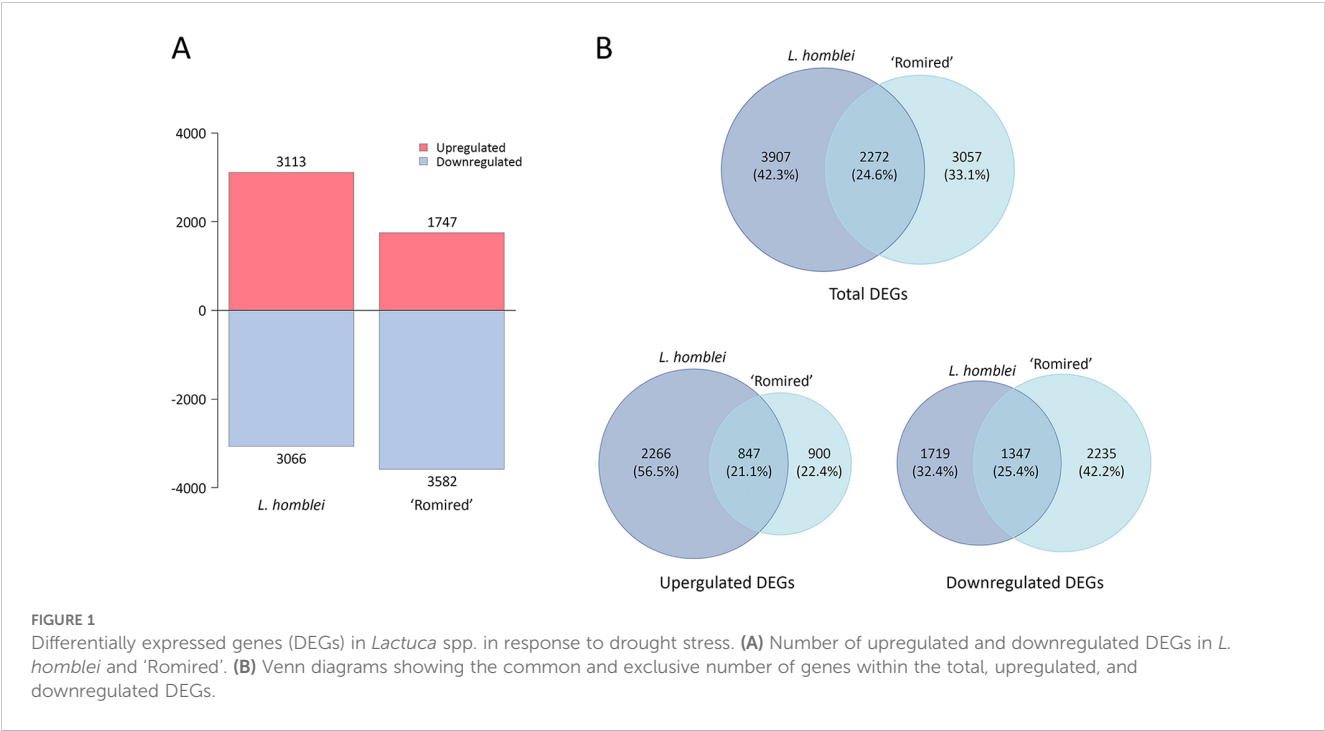
Group ^a	Sample	Raw reads	Clean reads	Mapped reads	Mapping rate (%)	GC (%)
<i>L. homblei</i> C	<i>L. homblei</i> 1	51,853,084	51,749,303	19,128,708	36.96	44.5
	<i>L. homblei</i> 2	46,166,391	46,074,390	17,230,542	37.40	45.5
	<i>L. homblei</i> 3	42,697,894	42,615,564	15,099,405	35.43	43.5
<i>L. homblei</i> DI	<i>L. homblei</i> 4	43,141,038	43,063,549	15,173,600	35.24	45.0
	<i>L. homblei</i> 5	40,836,180	40,756,877	13,200,520	32.39	45.0
	<i>L. homblei</i> 6	46,624,421	46,532,339	16,185,998	34.78	44.0
'Romired' C	'Romired' 1	35,523,451	35,428,117	29,244,585	82.55	41.5
	'Romired' 2	101,124,691	100,936,321	85,092,426	84.30	44.5
	'Romired' 3	110,827,310	110,624,887	93,466,553	84.49	45.0
'Romired' DI	'Romired' 4	94,626,862	94,435,273	80,092,100	84.81	45.0
	'Romired' 5	52,365,850	52,260,702	42,353,540	81.04	45.0
	'Romired' 6	41,007,770	40,932,006	34,429,220	84.11	45.0

^aC, control; DI, deficit irrigation.

upregulated and 3,582 downregulated genes) (Figure 1A). A total of 2,272 DEGs were common to both accessions: 847 genes were upregulated, 1,347 downregulated, and 78 exhibited an opposite behaviour in the two accessions (Figure 1B). Attending to the differences, the CWR showed a total number of DEGs higher than the cultivated species (42.3% of exclusive DEGs in *L. homblei* vs. 33.1% in 'Romired'). The same happened in the case of the upregulated genes, where the disparity was the largest, a 56.5% of the DEGs was exclusively upregulated in *L. homblei*, which was more than twice the upregulated DEGs only in 'Romired' (22.4%). In the case of the downregulated genes, we observed the

opposite, the number was higher in the cultivated species than in the wild relative (42.2% vs. 32.4%, respectively) (Figure 1B).

To deeply explore the DEG functions in the drought response of *Lactuca* spp., analyses of GO enrichment were conducted using the GO annotations found in Reyes-Chin-Wo et al. (2017). The three main GO categories, biological processes, cellular components and molecular functions, were studied within the upregulated and downregulated genes (Figure 2). Within the upregulated genes, the number of enriched GO terms in biological processes was higher in *L. homblei* than in 'Romired' (Figure 2A). In particular, an important number of *L. homblei* DEGs belonged to significantly



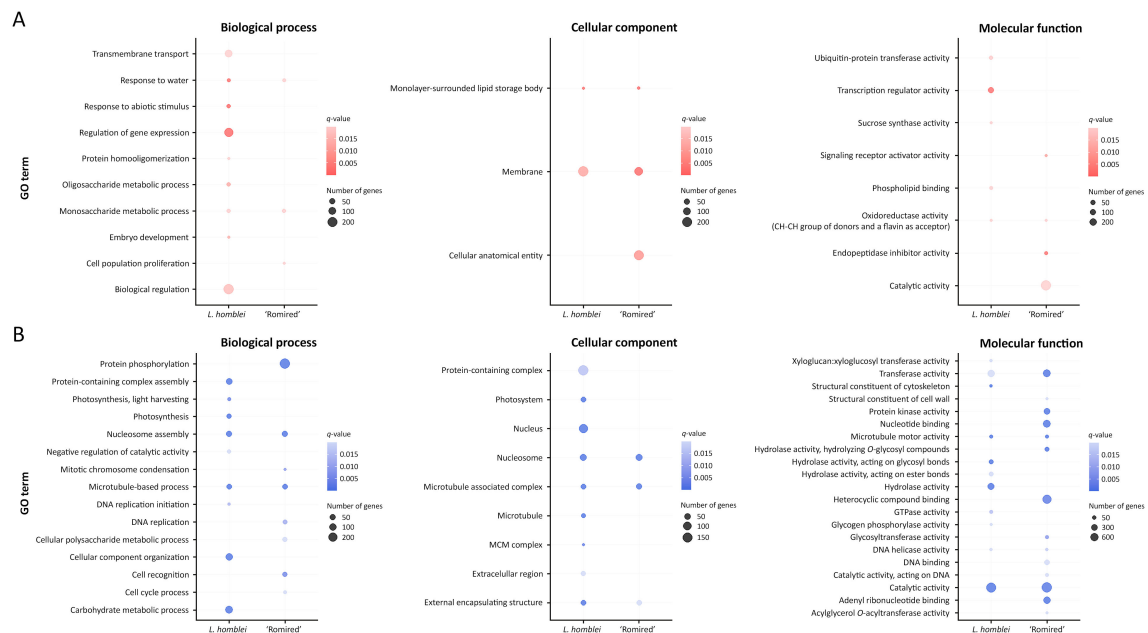


FIGURE 2

Enriched GO terms of *L. homblei* and 'Romired' within the (A) upregulated and (B) downregulated differentially expressed genes (DEGs) identified in a drought stress experiment for the three main GO categories: biological process, cellular component, and molecular function.

enriched GOs that were involved in transmembrane transport and different metabolic processes, though the most significantly upregulated DEGs were those in enriched GOs related to gene expression regulation and response to abiotic stimulus, water included. The response to water resulted to be also among the enriched GO terms in 'Romired', but the number of genes, and especially the significance level, were lower than in the CWR *L. homblei*. In the case of cellular components (Figure 2A), a similar number of genes were part of the GO terms membranes and lipid storage bodies in both species, with a higher significance in 'Romired'. In addition, the GO term cellular anatomical entity was enriched exclusively in the cultivated accession. Finally, for the molecular function category (Figure 2A), the most significantly enriched term was the transcription regulator activity in the CWR *L. homblei*, and the endopeptidase inhibitor activity in the commercial lettuce 'Romired'. Interestingly, anthocyanins (among other flavins) could be involved in the oxidoreductase activity in which a flavin group acts as acceptor, being actually the only enriched GO common to both species.

Within the downregulated DEGs of biological processes (Figure 2B), genes involved in the carbohydrate metabolism were the most represented in *L. homblei*, while in 'Romired' were those implied in protein phosphorylation, that in fact, appeared only in this accession. However, several enriched processes, as well as their significance levels, were common or similar in both species, such as those related to cellular division and multiplication (DNA replication, nucleosome assembly, and microtubule-based processes). In the cellular component category (Figure 2B), the enriched GO terms found in 'Romired' appeared also enriched in *L. homblei*: nucleosome, microtubule associated complex, and external encapsulating structure, with similar significance and number of

genes, except for the external encapsulating structure that resulted more significant in *L. homblei*. In fact, many more GO terms were enriched in *L. homblei*, with the nucleus and the protein-containing complex being the most represented ones. On the contrary, we found many more downregulated DEGs with enriched GO, as well as more terms and with a higher significance, in the cultivated ('Romired') than in the wild species (*L. homblei*) in the molecular function category (Figure 2B).

Some DEGs were assigned to more than one GO term, either because a gene can participate in different biological processes and molecular functions and be part of different cellular components or because GO is loosely hierarchical, with genes belonging to both 'parent' and 'child' terms. Thus, counting genes only once within each category, we obtained that the number of DEGs with enriched GO terms was very similar between *L. homblei* and 'Romired' within the upregulated genes (746 vs. 794, respectively), while it was lower in the CWR *L. homblei* than in the commercial variety 'Romired' within the downregulated ones (1,591 vs. 1978, respectively). Even so, the percentages of common DEGs in the two species was considerably lower in the case of upregulated genes than in downregulated: 3.23%, 18.25%, and 1.14% vs. 12.81%, 22.98%, and 23.00% in biological processes, cellular components and molecular functions, respectively.

Two heatmaps constructed using normalised expression data of DEGs in *L. homblei* and 'Romired' confirmed the effect of the drought stress treatment in *Lactuca* spp. plants in terms of gene regulation (Figure 3). A hierarchical clustering conducted with all the DEGs allowed us to identify two separate groups in the two accessions, as expected, the upregulated and the downregulated ones. This clustering also divided clearly the two conditions (C and DI) in both species, what was even more evident for *L. homblei*

(Figure 3A). In addition, both heatmaps showed again that the number of upregulated genes under water deficit was clearly higher in the CWR *L. homblei* (Figure 3A), whereas those downregulated were more numerous in the commercial variety 'Romired' (Figure 3B). These results show that the wild species was activating more mechanisms in response to drought stress. Figure 3 also shows two heatmaps constructed using the data of the anthocyanin content variation as a consequence of the drought stress for *L. homblei* and 'Romired' (Medina-Lozano et al., 2024), and the data of the two treatments themselves (C and DI). Content of all detected anthocyanins was higher under DI treatment than in C conditions in both *Lactuca* spp. Similar to what happened with the upregulated genes, the accumulation of anthocyanins in response to water stress was higher (and only significant) in *L. homblei* (Figure 3A). Actually, in *L. homblei* all the DI replicates showed a higher content than the C replicates. This was especially remarkable in the case of peonidin 3-O-glucoside, which was present under DI conditions and in the 3 biological replicates, but not under C conditions (Figure 3A). However, differences were not so clear (and not significant) between DI and C replicates in 'Romired', despite mean anthocyanin content being higher in DI than in C conditions, as commented above (Figure 3B). Even though, one of the anthocyanins was also identified exclusively under DI conditions in 'Romired', as observed in *L. homblei*, but in this case, it was cyanidin 3-(6"-acetylglucoside) (Figure 3B).

Nevertheless, the molecular mechanisms underlying the anthocyanin accumulation as a consequence of water deficiency have been barely studied in lettuce, unlike in other crops like grapevine (Castellarin et al., 2007) or canola (Chen et al., 2022b). To gain a more comprehensive understanding of the process in *Lactuca* spp., we mapped the expression profiles of the DEGs identified in the drought experiment which participate in the biosynthesis pathway of the detected anthocyanins (the general phenylpropanoid pathway and the flavonoid pathway, this last one

leading specifically to the anthocyanin biosynthesis) (Figure 4). We found a total of 36 DEGs involved in the pathway in either *L. homblei*, 'Romired' or both. Different expression profiles were observed between both species. Our results confirmed that, in these routes, more DEGs were activated in the CWR *L. homblei* than in the commercial variety 'Romired', 17 vs. 7 upregulated genes, respectively, what was concordant with the higher accumulation of anthocyanins in the wild relative (Medina-Lozano et al., 2024). Not all the isoforms of the genes coding for the enzymes catalysing each step were upregulated under DI. The activation happened mainly in the first steps of the pathway, that is, in the early biosynthesis genes (EBGs), especially at the beginning of anthocyanin-specific route (flavonoid pathway). This becomes glaringly obvious in the first step which is catalysed by the chalcone synthase (CHS), whose gene isoforms are all strongly and significantly upregulated only in *L. homblei* (triggering of the anthocyanin synthesis in the wild species). This pattern is not so obvious in the preceding genes from the general phenylpropanoid pathway as they participate in the biosynthesis of many other compounds apart from anthocyanins. Furthermore, the late biosynthesis genes (LBGs) were mostly upregulated in the CWR *L. homblei* but not in all the isoforms as observed in CHS (EBG), except in the last step which leads to the synthesis of the specific major anthocyanin (cyanidin 3-O-(6'-O-malonylglucoside)) where most of the genes coding for the isoforms were significantly activated in *L. homblei*. The final steps to produce the two minor anthocyanins, cyanidin 3-(6"-acetylglucoside) and peonidin 3-O-glucoside, are not clearly described in the literature. They might be catalysed by some acetyltransferases and O-methyltransferases, respectively, as suggested by Ino and Yamaguchi (1993) and Huguene et al. (2009), respectively. It is possible that the genes coding for these enzymes were activated under drought as those anthocyanins were detected in 'Romired' and *L. homblei*, respectively, only under stress conditions though they have not been characterised in *L. sativa* yet.

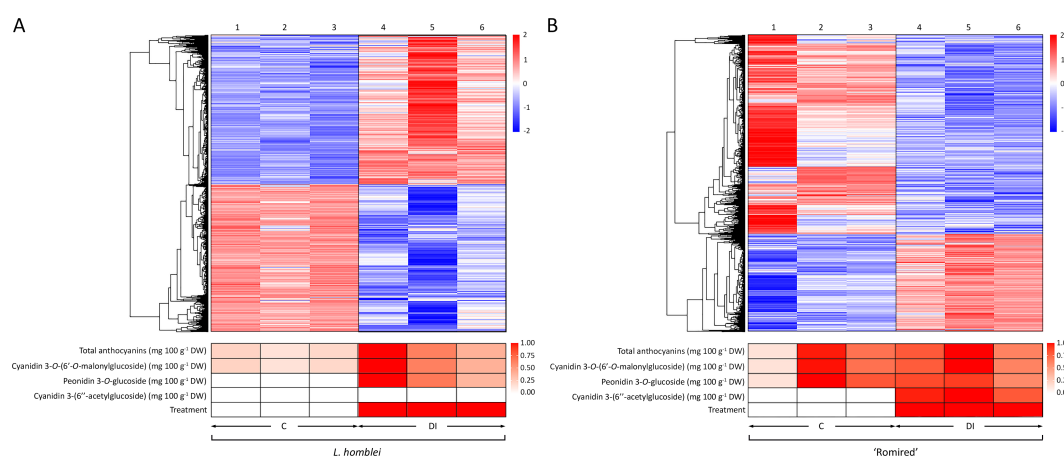


FIGURE 3

Heatmap representation of hierarchical analysis of the expression data from the differentially expressed genes (DEGs), as well as of total and individual anthocyanin content and treatment in (A) *L. homblei* and (B) 'Romired' under control (C) and deficit irrigation (DI) conditions. Numbers 1-3 and 4-6 show the biological replicates under C and DI, respectively. Phenotypic heatmaps represent scaled data from 0.2 to 1 within each compound, except for minor anthocyanins (peonidin 3-O-glucoside in *L. homblei* and cyanidin 3-(6"-acetylglucoside) in 'Romired') where data, as well as treatments, were scaled from 0 to 1.

3.3 Selection of candidate genes among the DEGs

Selection of anthocyanin-related genes potentially involved in the response to drought stress was based on different criteria. First, we searched for important changes in the expression levels. Second, we selected DEGs with high (positive and negative) and significant values of correlation with both anthocyanin content and drought stress treatment, obtained through a WGCNA. WGCNA allows to identify genes correlated with certain traits (anthocyanins and irrigation treatment in our case) to reveal putative genes with particular interest (Horvath and Dong, 2008). Lastly, we paid attention to gene function, so that DEGs were related to stress and/or anthocyanin content. Finally, 19 genes were selected for validation through real-time qPCR. Remarkably, genes meeting all these criteria resulted to be differentially expressed exclusively in *L. homblei*.

L. homblei $|\log_2(\text{FC})|$ values between C and DI ranged from 4.03 to 6.23 (Table 2). Both up- and downregulated genes were included in the selection. The higher accumulation of anthocyanins and/or the activation of stress response may result either from the upregulation of activators or from the downregulation of repressors. This was also observed in a previous study that characterised four genes related to anthocyanin content in lettuce (Su et al., 2020). From the WGCNA results, we obtained absolute correlation values with anthocyanins ranging from 0.81 to 0.93, and with treatment, from 0.86 to 0.99 (Table 2). Both positive and negative correlations were also considered here. By contrast, in the case of 'Romired', we found 5 out of the 19 genes showing a significant correlation with the anthocyanin content, but none of them exhibited a significant change of expression level nor a significant correlation with the stress treatment (data not shown).

3.4 Validations of candidate genes by qPCR

The expression data of the 19 selected genes obtained from the RNA-seq analysis were validated by real-time qPCR. In *L. homblei*, the 13 downregulated and the 6 upregulated genes according to the RNA-seq analysis showed concordant expression profiles with the qPCR results (Figure 5). Significant, and very significant differences were observed between C and DI treatments for seven and two genes, respectively, according to qPCR data (Figure 5). In the case of 'Romired', the selected genes did not show any differential expression in the RNA-seq analysis, as mentioned above, nor by qPCR. Even so, the expression of 15 out of the 19 genes followed the same profile using the two different techniques (Supplementary Figure S1). Therefore, we were able to confirm the reliability of the results from the RNA-seq analysis.

The expression profiles of the selected DEGs obtained with both techniques (RNA-seq and qPCR) were also analysed by hierarchical clustering both using the mean values (Figure 6) and all data (Supplementary Figure S2). According to the RNA-seq data, two clearly differentiated expression patterns could be observed for *L. homblei*: the expression levels were noticeably lower in C than in DI conditions in the case of upregulated genes, and vice versa in the

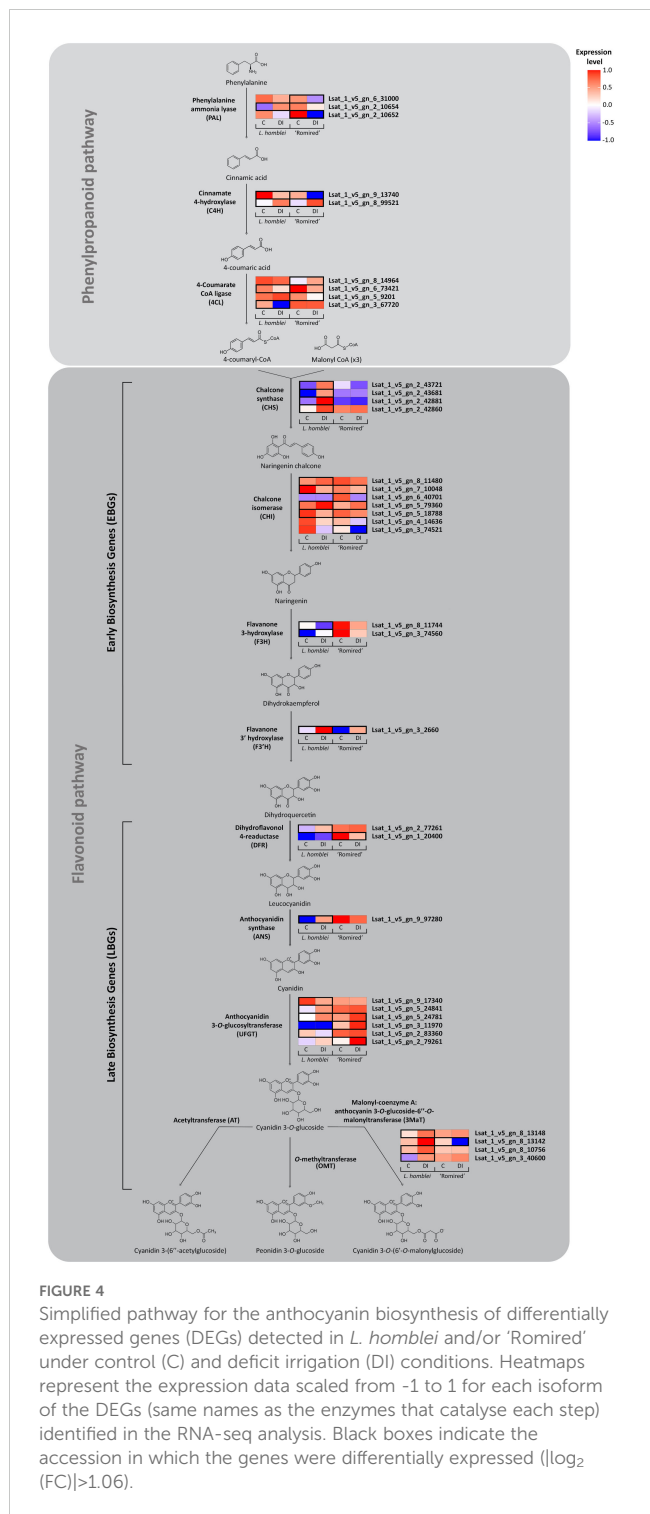


FIGURE 4

Simplified pathway for the anthocyanin biosynthesis of differentially expressed genes (DEGs) detected in *L. homblei* and/or 'Romired' under control (C) and deficit irrigation (DI) conditions. Heatmaps represent the expression data scaled from -1 to 1 for each isoform of the DEGs (same names as the enzymes that catalyse each step) identified in the RNA-seq analysis. Black boxes indicate the accession in which the genes were differentially expressed ($|\log_2(\text{FC})| > 1.06$).

case of downregulated genes (Figure 6A; Supplementary Figure S2A). In contrast, for 'Romired' the differences in expression were not clear in most of the selected genes (Figure 6A; Supplementary Figure S2A), as was expected since they did not result to be differentially expressed in the RNA-seq analysis. Comparable patterns, especially for *L. homblei* samples, were observed when qPCR data were represented (Figure 6B; Supplementary Figure S2B).

TABLE 2 Gene product, regulation and correlation with total anthocyanin content and treatment of the 19 differentially expressed genes (DEGs) selected.

Gene ID	Gene product	Regulation	Log ₂ (FC) ^a	FDR ^b	Total anthocyanins		Irrigation treatment	
					Correlation coefficient	p value	Correlation coefficient	p value
Lsat_1_v5_gn_1_21441	Subtilisin-like protease SBT3	Downregulated	-4.03	3.33E-04	-0.81	0.05	-0.96	2.85E-03
Lsat_1_v5_gn_1_50480	Haloacid dehalogenase (HAD)-like hydrolase superfamily protein	Downregulated	-4.28	3.33E-04	-0.93	0.01	-0.86	0.03
Lsat_1_v5_gn_1_109200	DNA damage-repair/toleration protein DRT100	Downregulated	-4.44	2.27E-03	-0.82	0.05	-0.91	0.01
Lsat_1_v5_gn_1_127541	14 kDa proline-rich protein DC2.15	Downregulated	-5.14	2.53E-03	-0.93	0.01	-0.91	0.01
Lsat_1_v5_gn_2_15680	GDSL esterase/lipase	Downregulated	-4.81	0.04	-0.93	0.01	-0.89	0.02
Lsat_1_v5_gn_2_43400	Probable pectate lyase 8	Downregulated	-5.52	2.25E-03	-0.87	0.02	-0.99	1.75E-04
Lsat_1_v5_gn_2_47181	Protein ECERIFERUM 26	Downregulated	-5.16	0.03	-0.92	0.01	-0.91	0.01
Lsat_1_v5_gn_2_90361	Gibberellin-regulated protein 6	Downregulated	-4.29	0.03	-0.93	0.01	-0.95	3.94E-03
Lsat_1_v5_gn_2_116640	Heat shock cognate 70 kDa protein 2	Upregulated	5.53	6.34E-06	0.84	0.03	0.97	1.06E-03
Lsat_1_v5_gn_3_1101	Zinc finger protein ZAT1	Downregulated	-4.75	1.31E-03	-0.90	0.01	-0.98	5.92E-04
Lsat_1_v5_gn_3_20640	PRA1 family protein E	Upregulated	4.18	1.18E-03	0.82	0.04	0.98	8.11E-04
Lsat_1_v5_gn_5_7401	NAC transcription factor 56	Upregulated	4.22	7.39E-04	0.82	0.05	0.94	0.01
Lsat_1_v5_gn_5_10141	Protein MHF1 homolog	Downregulated	-4.66	0.04	-0.91	0.01	-0.86	0.03
Lsat_1_v5_gn_5_26000	B-box zinc finger protein 21-like	Upregulated	4.34	2.38E-05	0.82	0.05	0.93	0.01
Lsat_1_v5_gn_6_67540	Type I inositol polyphosphate 5-phosphatase 2	Downregulated	-4.32	7.53E-07	-0.85	0.03	-0.93	0.01
Lsat_1_v5_gn_7_92980	Ribonuclease III-like protein RTL3	Downregulated	-6.23	3.97E-03	-0.92	0.01	-0.93	0.01
Lsat_1_v5_gn_8_157561	Transcription factor MYC/MYB N-terminal domain-containing protein	Downregulated	-4.46	4.60E-04	-0.84	0.04	-0.94	4.91E-03
Lsat_1_v5_gn_8_165301	Phospholipase A1 phospholipid-inositol phosphatase PLIP2	Upregulated	4.28	4.73E-05	0.84	0.04	0.96	1.94E-03
Lsat_1_v5_gn_9_80621	Amino acid permease 6	Upregulated	4.17	2.91E-03	0.82	0.04	0.96	2.81E-03

^aFC, fold change.
^bFDR, False Discovery Rate.

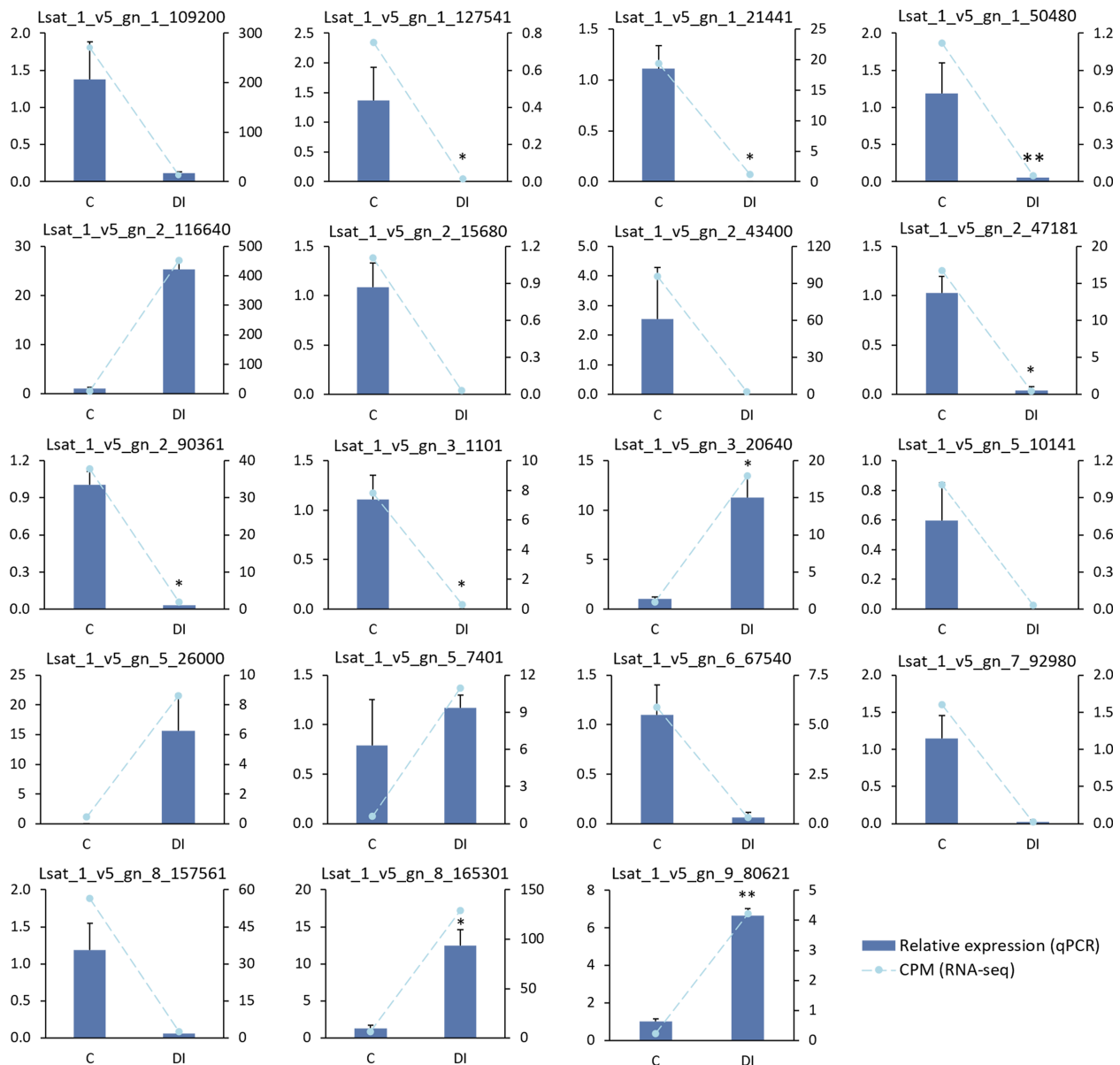


FIGURE 5

Expression data obtained by qPCR (relative expression) and by RNA-seq (CPM, counts per million) of 19 selected genes in the wild species *L. homblei* under control (C) and deficit irrigation (DI) conditions. Bars represent standard error of the total ($n=3$). * $p<0.05$, ** $p<0.01$. Transformations were applied to achieve normal distribution in qPCR data in the following cases: $1/(1+x)^2$ to Lsat_1_v5_gn_1_109200, Lsat_1_v5_gn_1_127541, Lsat_1_v5_gn_1_50480, Lsat_1_v5_gn_2_116640, Lsat_1_v5_gn_2_43400, and Lsat_1_v5_gn_3_20640; and $1/\sqrt{(x+1)}$ to Lsat_1_v5_gn_1_21441. Wilcoxon test was used with non-normally distributed qPCR data of Lsat_1_v5_gn_5_10141 and Lsat_1_v5_gn_5_26000.

3.5 Putative function of validated candidate genes

Turning the attention to gene function, it could be confirmed that those 19 DEGs with large changes in expression levels and high correlations with treatment and anthocyanins were indeed related to stress responses and/or to anthocyanin content (Table 3). Specifically, most gene products of the selected DEGs have been described to participate in the response to one or more types of stresses. Several of the DEGs are involved in the response to biotic stresses, like resistance to bacteria (Lsat_1_v5_gn_1_21441 (Ramírez et al., 2013)), virus (Lsat_1_v5_gn_3_1101 (Tsitssekian et al., 2023)), or

fungi (Lsat_1_v5_gn_3_20640 (Wu et al., 2022)), but most have been described to act in abiotic stress responses. In particular, genes related to water deficit and/or the stress-responsive hormone ABA (abscisic acid) stood out, such as Lsat_1_v5_gn_2_116640 (Clément et al., 2011), Lsat_1_v5_gn_2_43400 (Palusa et al., 2007), and Lsat_1_v5_gn_3_20640 (Tahmasebi et al., 2019). Furthermore, not only genes reported to be generally activated under water stress conditions were included in the selection, but also some described as negative regulators, which were in fact downregulated (inhibition of suppressors) in our samples subject to the drought treatment (Tables 2, 3), like Lsat_1_v5_gn_1_50480 (Lee et al., 2022), Lsat_1_v5_gn_2_90361 (Qu et al., 2016), and

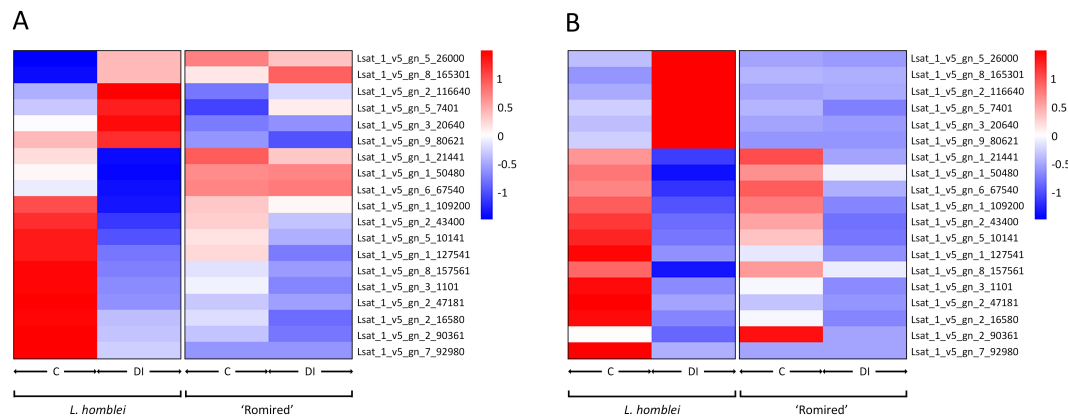


FIGURE 6

Heatmap representation of hierarchical analysis of the mean expression data ($n=3$) of 19 selected genes in *L. homblei* and 'Romired' under control (C) and deficit irrigation (DI) conditions according to (A) RNA-seq and (B) real-time qPCR analyses.

Lsat_1_v5_gn_6_67540 (Na and Metzger, 2020). Some of the selected DEGs have been related to other abiotic stresses such as salt (Lsat_1_v5_gn_1_127541 (Bhattarai et al., 2021) and Lsat_1_v5_gn_3_1101 (He et al., 2020)), heat (Lsat_1_v5_gn_2_47181 (Zhang et al., 2022)), and nutrient deficiency (Lsat_1_v5_gn_9_80621 (Zhou et al., 2021)), which makes sense as especially salt and heat stresses often occur simultaneously with drought.

Genes related to flavonoid (e.g., anthocyanins) accumulation were mostly upregulated in our samples (Tables 2 and 3). Two of those genes encode putative transcription factors (TFs): a zinc finger protein (Lsat_1_v5_gn_5_26000 (Zhang et al., 2021)) and a NAC TF (Lsat_1_v5_gn_5_7401 (Wei et al., 2020)), both having been described to induce anthocyanin-related genes. Among the DEGs identified in this work, other two have been previously described to cause the increase of flavonoid or anthocyanin content under abiotic stress. Specifically, Lsat_1_v5_gn_2_15680 was found to play important roles in flavonoid accumulation under drought stress in tea (Li et al., 2020), and Lsat_1_v5_gn_8_165301 was involved in the increase of anthocyanin content when overexpressed in *Arabidopsis thaliana* under ABA-mediated abiotic stress responses (Wang et al., 2018). Another selected gene (Lsat_1_v5_gn_8_157561) is not well characterised but could possibly be related to the anthocyanin content as it contains a MYB domain and, in plants, MYB TFs have been described as one of the major transcriptional regulators of anthocyanin pathway, both activators and repressors (Yan et al., 2021; Cao et al., 2024). This specific gene was downregulated in our samples when exposed to water stress, so it might be a transcriptional repressor of anthocyanin biosynthesis. Lsat_1_v5_gn_7_92980 encodes a ribonuclease III-like protein 3 (RTL3) that cleaves double-stranded RNA (Comella et al., 2008) and might be also related to anthocyanin content. Proteins with ribonuclease III domains have been described to participate in the regulation of seed coat in soybean and fruit colour in peach through the production of siRNAs (small interfering RNAs) and the increase of transcription levels of genes implied in anthocyanin regulation (Zhu et al., 2012; Jia et al., 2020). Finally, two genes that encode proteins involved in DNA damage repair were also selected

(Lsat_1_v5_gn_1_109200 (Fujimori et al., 2014), and Lsat_1_v5_gn_5_10141 (Dangel et al., 2014)). This is not surprising as the generation of ROS is a potential cause of DNA damage under drought stress and a fine-tuned regulation of DNA repair is required to tolerate it (Shim et al., 2018).

3.6 Polymorphisms in the DEGs

In silico search and prediction of polymorphisms were performed to get an overview of the variation in the sequences of the total number of DEGs detected and to explore more deeply the structural variation of the set of 19 selected DEGs.

A total of 235,600 polymorphisms were found in the whole set of DEGs (9,236) in *L. homblei* and 'Romired' compared to the reference genome (both shared and species-exclusive). Most polymorphisms were detected in *L. homblei*, as expected since it is a wild species that is very distant from the cultivated *L. sativa* used as reference. The predominant types of polymorphisms were SNPs (Single Nucleotide Polymorphisms) (89.22%), followed by MNPs (Multiple Nucleotide Polymorphisms) (9.97%), and, in a much smaller extent, by indels (insertions-deletions) (0.81%). The most abundant polymorphism effects were synonymous (63.20%) and missense (27.96%). We also identified intron (6.14%) and splice region (2.23%) variants, as well as others that were present in less than 0.1%, so they are not detailed here.

In the subset of 19 DEGs selected, a total of 404 polymorphisms with 408 predicted effects were identified in *L. homblei* (Table 4), in contrast to the 11 polymorphisms with 12 predicted effects found in those same 19 genes non-differentially expressed in 'Romired' (Supplementary Table S2). Considering only the 19 DEGs in *L. homblei*, the proportions of both polymorphism types and effects were almost the same than those found in the whole set of DEGs, 87.87% of polymorphisms were SNPs, 11.14% were MNPs, and 0.99% were indels. Once again, we found that the predominant predicted effect was synonymous (70.10%), followed by missense (27.44%) type (Table 4). We also identified, though in a reduced number of genes, effects in splice regions (0.98%) and introns (0.49%), disruptive and

TABLE 3 Putative function of the 19 candidate differentially expressed genes (DEGs) in *L. homblei*.

Gene ID	Gene product	Putative function	Reference
Lsat_1_v5_gn_1_21441	Subtilisin-like protease SBT3	Plant immune priming in systemic induced resistance establishment	Ramírez et al., 2013
Lsat_1_v5_gn_1_50480	Haloacid dehalogenase (HAD)-like hydrolase superfamily protein	Repression of ABA-response and ABA-mediated drought tolerance	Lee et al., 2022
Lsat_1_v5_gn_1_109200	DNA damage-repair/toleration protein DRT100	Repair and toleration of UV-B-induced DNA damage	Fujimori et al., 2014
Lsat_1_v5_gn_1_127541	14 kDa proline-rich protein DC2.15	Cell wall modification and organization	Bhattarai et al., 2021
Lsat_1_v5_gn_2_15680	GDSL esterase/lipase	Flavonoid accumulation and lipid reduction under drought stress	Li et al., 2020
Lsat_1_v5_gn_2_43400	Probable pectate lyase 8	Response to stimulus through cell wall modification	Palusa et al., 2007
Lsat_1_v5_gn_2_47181	Protein ECERIFERUM 26	Dehydration tolerance under heat stress	Zhang et al., 2022
Lsat_1_v5_gn_2_90361	Gibberellin-regulated protein 6	ABA-repressible peptide hormone precursor	Qu et al., 2016
Lsat_1_v5_gn_2_116640	Heat shock cognate 70 kDa protein 2	ABA-induced stomatal closure	Clément et al., 2011
Lsat_1_v5_gn_3_1101	Zinc finger protein ZAT1	Putative transcription factor that acts in the response to abiotic and biotic stresses	He et al., 2020 ; Tsitsekan et al., 2023
Lsat_1_v5_gn_3_20640	PRA1 family protein E	Protein transporter involved in abiotic and biotic stress responses	Tahmasebi et al., 2019 ; Wu et al., 2022
Lsat_1_v5_gn_5_7401	NAC transcription factor 56	Transcription factor that induces anthocyanin accumulation	Wei et al., 2020
Lsat_1_v5_gn_5_10141	Protein MHF1 homolog	DNA repair and homologous recombination	Dangel et al., 2014
Lsat_1_v5_gn_5_26000	B-box zinc finger protein 21-like	Positive transcriptional regulator of light-induced anthocyanin accumulation	Zhang et al., 2021
Lsat_1_v5_gn_6_67540	Type I inositol polyphosphate 5-phosphatase 2	Putative repressor of water stress response	Na and Metzger, 2020
Lsat_1_v5_gn_7_92980	Ribonuclease III-like protein RTL3	Cleavage of double strand RNA	Comella et al., 2008
Lsat_1_v5_gn_8_157561	Transcription factor MYC/MYB N-terminal domain-containing protein	Putative transcriptional repressor of anthocyanin biosynthesis	Yan et al., 2021
Lsat_1_v5_gn_8_165301	Phospholipase A1 phospholipid-inositol phosphatase PLIP2	ABA-mediated abiotic stress responses and anthocyanin accumulation	Wang et al., 2018
Lsat_1_v5_gn_9_80621	Amino acid permease 6	AA transport under nutrient stresses	Zhou et al., 2021

TABLE 4 Predicted effects for the polymorphisms detected in the 19 candidate differentially expressed genes (DEGs) in *L. homblei*.

Gene ID	Conservative in-frame deletion ^a	Disruptive in-frame deletion ^a	Frameshift variant ^a	Intron variant	Missense variant ^a	Splice region variant	Synonymous variant
Lsat_1_v5_gn_1_21441	–	–	–	–	23	–	27
Lsat_1_v5_gn_1_50480	–	–	–	–	3	–	5
Lsat_1_v5_gn_1_109200	–	–	–	–	4	–	27
Lsat_1_v5_gn_1_127541	–	–	–	–	1	–	3
Lsat_1_v5_gn_2_15680	–	–	–	–	3	–	3
Lsat_1_v5_gn_2_43400	–	–	–	–	5	1	31
Lsat_1_v5_gn_2_47181	–	–	–	–	7	–	10
Lsat_1_v5_gn_2_90361	–	–	–	–	2	–	3
Lsat_1_v5_gn_2_116640	–	1	–	2	4	–	51
Lsat_1_v5_gn_3_1101	1	–	1	–	12	–	8
Lsat_1_v5_gn_3_20640	–	–	–	–	4	–	11
Lsat_1_v5_gn_5_7401	–	–	–	–	7	–	16
Lsat_1_v5_gn_5_10141	–	–	–	–	1	1	4
Lsat_1_v5_gn_5_26000	–	–	–	–	8	–	12
Lsat_1_v5_gn_6_67540	–	–	–	–	5	–	11
Lsat_1_v5_gn_7_92980	–	–	–	–	3	–	2
Lsat_1_v5_gn_8_157561	–	–	–	–	3	2	24
Lsat_1_v5_gn_8_165301	–	1	–	–	13	–	19
Lsat_1_v5_gn_9_80621	–	–	–	–	4	–	19
Percentage	0.25	0.49	0.25	0.49	27.44	0.98	70.10

^aHigh and moderate effects are shown in bold.

conservative in-frame deletions (0.49 and 0.25%, respectively), and a frameshift variant (0.25%). The impact of the polymorphisms was frequently low, which makes sense considering that most of them were predicted to have a synonymous effect. However, a polymorphism with high impact was detected. It was a 2-bp insertion that theoretically causes a frameshift mutation in the Lsat_1_v5_gn_3_1101 gene of the wild species which is responsible for the appearance of a premature stop codon. A conservative in-frame deletion was also found in this same gene. According to our results, this gene was downregulated in *L. homblei* whereas in ‘Romired’ was not differentially expressed, in which showed low expression levels in both C and DI conditions. This gene codes for a zinc-finger protein and appears in the literature as a putative TF that intervenes in the response to abiotic stress (He et al., 2020). A possible effect of one or both polymorphisms might be that the truncated protein acts as a repressor in the wild species under C conditions but stops inhibiting its target(s) as a consequence of its own downregulation under water stress. Other polymorphisms with possible important effects were the putative disruptive in-frame deletions found in Lsat_1_v5_gn_2_116640 and in Lsat_1_v5_gn_8_165301 genes, whose predicted impact was moderate. Lsat_1_v5_gn_2_116640 encodes a 70-kDa heat shock

cognate protein. Heat shock proteins (HSPs) were initially described in relation to heat tolerance (Ritossa, 1962), although nowadays they are well known to be expressed in response to a great diversity of environmental stressors besides heat (reviewed in Ul Haq et al. (2019)). According to the RNA-seq analysis, this gene (Lsat_1_v5_gn_2_116640) showed a considerable increase in expression in *L. homblei* in response to drought, whereas in ‘Romired’ there was no significant change, with the values under both C and DI being similar to those in C plants of *L. homblei*. Therefore, the disruptive in-frame deletion in this gene could be inducing the activation of this HSP when *L. homblei* plants are subject to drought stress. Lsat_1_v5_gn_8_165301 encodes a Phospholipase A1 phospholipid-inositol phosphatase 2 (PLIP2) that has been described to be involved in the accumulation of anthocyanins under ABA-mediated abiotic stress responses (Wang et al., 2018). This gene also exhibited a highly significant upregulation in *L. homblei* and no change of expression in ‘Romired’. In this case, its expression levels in ‘Romired’ under C and DI conditions were similar to those of *L. homblei* under DI. Thus, the disruptive in-frame deletion found in *L. homblei* sequence might be causing the gene to be activated only under stress in the wild plants.

4 Discussion

4.1 Identification and analysis of DEGs under drought stress conditions

The number of exclusive up- and downregulated genes was more than twice in *L. homblei* and 1.3 times higher in 'Romired', respectively. In general, genes related to regulation within biological process (e.g., response to water) and molecular function categories were more abundant and more intensively upregulated in the wild species whereas genes responsible for cellular components were more commonly and significantly upregulated in the cultivated species. In the case of the downregulated DEGs, the most represented terms in both species were those related to catalytic activities. In general terms, basal and growth-related processes were deactivated in both species, which probably contributes to redirect resources to guarantee plant survival.

Interestingly, activation of responses seemed to be species specific, and it looks like the CWR was triggering more mechanisms of response to drought stress as the number of upregulated genes was clearly larger in *L. homblei* under DI. In contrast, the higher number of downregulated DEGs common to both species, many of them implied in basal processes, could be due to the deactivation of basal metabolism processes to designate more resources to water deficit tolerance, previously described in different plant species subject to water stress (Shao et al., 2009). The results from the GO enrichment analysis are in agreement with those found in other studies that assessed different stresses in lettuce, in which response to stimulus, biological regulation, metabolic processes, binding and catalytic activities, as well as membrane components, were the most represented terms (Wang et al., 2017; Zhou et al., 2023). Interestingly, in a transcriptomic analysis carried out to identify genes involved in lettuce anthocyanin accumulation, the most represented GO terms were the same (Zhang et al., 2016).

Both anthocyanin contents and the number of upregulated genes were clearly larger in *L. homblei* under DI. Therefore, results point to a relationship between gene expression profiles (for some DEGs) and changes in the accumulation of these antioxidant compounds. In addition, not only the most abundant anthocyanins showing the biggest change in quantity in response to water stress but also the minor anthocyanins only identified under DI (peonidin 3-*O*-glucoside and cyanidin 3-(6''-acetylglucoside) in *L. homblei* and 'Romired', respectively) could play a role in the response to drought. That is more plausible in the case of *L. homblei* where the differences in anthocyanin content between C and DI conditions were significant (Medina-Lozano et al., 2024).

Interestingly, to activate the anthocyanin biosynthesis route in *L. homblei* seems to be enough to upregulate the isoforms of the gene controlling the first step of the specific pathway branch (i.e., *CHS*), even when the preceding genes from the general phenylpropanoid pathway (e.g., 4-coumarate-CoA ligase (*4CL*)) could be downregulated or not significantly differentially expressed, as they are involved in the biosynthesis of many other phenylpropanoids apart from anthocyanins. Something similar has

been shown in a previous study on the expression of those genes and the anthocyanin content of poplar leaves (Tian et al., 2021).

All these differences between both *Lactuca* spp. might reflect a larger plasticity of the wild species to adapt to environmental changes. The great genetic diversity of wild species allows them to counteract the effects of different stresses more effectively (Jordanovska et al., 2020), whereas the cultivated species could have lost these mechanisms through domestication. In fact, the common DEGs to both accessions which show an opposite sense in the change of expression could consist of genes that have acquired a different mode of action as *L. homblei* belongs to the lettuce tertiary gene pool, the most genetically distant from *L. sativa*. Alternatively, they could be artefacts, either methodological (e.g., library preparation) or statistical or even both.

4.2 Validation, putative function, and polymorphisms of candidate DEGs

The fact that the genes with the strongest change of expression and correlation with anthocyanin content and drought were differentially expressed only in *L. homblei* might reveal, once again, the existence of tolerance mechanisms in the wild species that are not present in the cultivated one. This is in agreement with the wild species showing the highest increase (and the only resulting statistically significant) of anthocyanins in a previous study on drought stress with the same accessions, among others (Medina-Lozano et al., 2024).

The reliability of the results from the RNA-seq analysis was confirmed as the candidate DEGs were validated by real-time qPCR, being all differentially expressed only in the wild species, in which the expression profiles obtained with the two techniques coincided. Besides, most gene products of the selected DEGs have been described to participate in the response to one or more types of stresses which makes sense as some stresses often occur simultaneously.

The fact that the genes with a high change in the level of expression identified in this study resulted to be related to both the stress response and the anthocyanin content could indicate that these compounds are playing an important role for plants to cope with the drought conditions, as was also proposed before in purple-stem *Brassica napus* L. (Chen et al., 2022b).

Talking about the polymorphisms found in all DEGs and in the candidate genes, our results are in agreement with studies carried out in other crops that also used transcriptomic data, in which the number of SNPs was also much higher than the number of indels, and the most abundant polymorphism effects were synonymous and missense variants too (Iquebal et al., 2017; Muñoz-Espinoza et al., 2020). The impact of the polymorphisms identified in the candidate DEGs frequently resulted low as most of them were predicted to have a synonymous effect. However, a few of them showed a high or moderate predicted impact what could be a reflection of the drastic changes in the gene expression profiles (either activation or inhibition) in the wild species when subject to

drought stress. Among the rest of the polymorphisms found, the missense variants could have an impact on the function of the resultant protein, though this one has been predicted to be moderate.

In silico tools are truly useful for obtaining information of functional and structural variants on the transcriptome and their possible correlation with phenotypic changes (Yazar and Özbek, 2021). However, further experimental approaches like functional analyses are essential to verify in the future the polymorphism effect found in putative candidate genes involved in the anthocyanin accumulation and the response to drought stress.

5 Conclusion

Mechanisms of response to drought stress related to anthocyanins were triggered in the wild species *L. homblei* but not in the cultivated lettuce variety 'Romired'. The involvement of the proposed candidate genes in the increase of anthocyanin content and the response to drought stress in the wild species is supported by their large and significant changes in the expression levels when the plants were subjected to water deprivation and by their high correlation with anthocyanin content. Furthermore, the activation of the anthocyanin biosynthesis route was mainly achieved by significantly upregulating the genes controlling the first step of the specific branch (flavonoid pathway), again exclusively in the wild species.

All the candidate genes have been reported before to be involved in the response to biotic or abiotic stresses in other species (but not in lettuce), what demonstrate that plants have developed interconnected and interacting routes to deploy integrated responses to combinations of concurrent stresses.

This wild species has become a potential donor of drought tolerance genes to the cultivated lettuce that foreseeably will make the crop more resilient and sustainable, while containing more beneficial compounds (i.e., anthocyanins) for human health.

Data availability statement

The datasets presented in this study can be found in the European Nucleotide Archive (ENA) at <https://www.ebi.ac.uk/ena/browser/home> accession number: PRJEB75159.

Author contributions

IML: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. JG: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Supervision, Writing – review &

editing. AD: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded by the projects PID2022-138484OR-I00 from the Spanish Ministry of Science and Innovation and State Research Agency (AEI) and LMP148_21 from the Government of Aragón; and by the Operational Programme FEDER Aragón 2023-2025 and 2020-2022, and the European Social Fund from the European Union (A12-23R: “Grupo de investigación en fruticultura: caracterización, adaptación y mejora genética”). IML was supported by a predoctoral contract for training doctors from the Spanish Ministry of Science, Innovation and Universities (MCIU) and the Spanish State Research Agency (AEI).

Acknowledgments

We thank J. A. Aranjuelo for technical support and D. L. Goodchild for reviewing the English language. We gratefully acknowledge the Centre for Genetic Resources (CGN, Wageningen, Netherlands) for supplying the seeds used here.

Conflict of interest

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

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The authors declare that Generative AI was not used in the creation of this manuscript.

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.2024.1494339/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Primer sequences, amplicon length, and annealing temperature of the reference gene (*TRXL3-3*) and the 19 differentially expressed genes (DEGs) selected to be validated by qPCR.

SUPPLEMENTARY TABLE 2

Predicted effects for the polymorphisms detected in the 19 candidate differentially expressed genes (DEGs) in 'Romired'.

References

- Becker, C., Klaering, H. P., Kroh, L. W., and Krumbein, A. (2014). Cool-cultivated red leaf lettuce accumulates cyanidin-3-O-(6"-O-malonyl)-glucoside and caffeoylmalic acid. *Food Chem.* 146, 404–411. doi: 10.1016/j.foodchem.2013.09.061
- Bhattarai, S., Fu, Y. B., Coulman, B., Tanino, K., Karunakaran, C., and Biligetu, B. (2021). Transcriptomic analysis of differentially expressed genes in leaves and roots of two alfalfa (*Medicago sativa* L.) cultivars with different salt tolerance. *BMC Plant Biol.* 21, 446. doi: 10.1186/s12870-021-03201-4
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Cao, Y., Mei, Y., Zhang, R., Zhong, Z., Yang, X., Xu, C., et al. (2024). Transcriptional regulation of flavonol biosynthesis in plants. *Hortic. Res.* 11, uhae043. doi: 10.1093/hr/uhae043
- Castellarin, S. D., Pfeiffer, A., Sivilotti, P., Degan, M., Peterlunger, E., and Di Gasparo, G. (2007). Transcriptional regulation of anthocyanin biosynthesis in ripening fruits of grapevine under seasonal water deficit. *Plant Cell Environ.* 30, 1381–1399. doi: 10.1111/j.1365-3040.2007.01716.x
- Chen, L., Xu, M., Liu, C., Hao, J., Fan, S., and Han, Y. (2022a). LsMYB15 regulates bolting in leaf lettuce (*Lactuca sativa* L.) under high-temperature stress. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.921021
- Chen, W., Miao, Y., Ayyaz, A., Hannan, F., Huang, Q., Ulhassan, Z., et al. (2022b). Purple stem *Brassica napus* exhibits higher photosynthetic efficiency, antioxidant potential and anthocyanin biosynthesis related genes expression against drought stress. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.936696
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms. *SnpEff. Fly* 6, 80–92. doi: 10.4161/fly.19695
- Clément, M., Leonhardt, N., Droillard, M.-J., Reiter, I., Montillet, J.-L., Genty, B., et al. (2011). The cytosolic/nuclear HSC70 and HSP90 molecular chaperones are important for stomatal closure and modulate abscisic acid-dependent physiological responses in *Arabidopsis*. *Plant Physiol.* 156, 1481–1492. doi: 10.1104/pp.111.174425
- Comella, P., Pontvianne, F., Lahmy, S., Vignols, F., Barbezier, N., DeBures, A., et al. (2008). Characterization of a ribonuclease III-like protein required for cleavage of the pre-rRNA in the 3' ETS in *Arabidopsis*. *Nucleic Acids Res.* 36, 1163–1175. doi: 10.1093/nar/gkm1130
- Dangel, N. J., Knoll, A., and Puchta, H. (2014). MHF1 plays Fanconi anaemia complementation group M protein (FANCM)-dependent and FANCM-independent roles in DNA repair and homologous recombination in plants. *Plant J.* 78, 822–833. doi: 10.1111/tpj.12507
- Eriksen, R. L., Knepper, C., Cahn, M. D., and Mou, B. (2016). Screening of lettuce germplasm for agronomic traits under low water conditions. *HortScience* 51, 669–679. doi: 10.21273/hortsci.51.6.669
- Everaert, C., Luypaert, M., Maag, J. L. V., Cheng, Q. X., Dinger, M. E., Hellemans, J., et al. (2017). Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. *Sci. Rep.* 7, 1559. doi: 10.1038/s41598-017-01617-3
- FAO. (2021). *The impact of disasters and crises on agriculture and food security: 2021* (Rome, Italy). doi: 10.4060/cb3673en
- FAOSTAT. (2021). *Statistics of the Food and Agriculture Organization of the United Nations*. Available online at: <http://www.fao.org/faostat/en/data/QC> (Accessed January 10, 2024).
- Faria, D. (2017). *GOEnrichment* (GitHub Repos). Available online at: <https://github.com/DanFaria/GOEnrichment> (Accessed February 10, 2024).
- Farooq, M., Wahid, A., Kobayashi, N., Fujita, D., and Basra, S. M. A. (2009). "Plant drought stress: effects, mechanisms and management," in *Sustainable Agriculture*. Eds. E. Lichtfouse, M. Navarrete, P. Debaeke, S. Véronique and C. Alberola (Springer, Dordrecht), 153–188. doi: 10.1007/978-90-481-2666-8_12
- Fujimori, N., Suzuki, N., Nakajima, Y., and Suzuki, S. (2014). Plant DNA-damage repair/tolerance 100 protein repairs UV-B-induced DNA damage. *DNA Repair.* 21, 171–176. doi: 10.1016/j.dnarep.2014.05.009
- Garcia, C., and Blesso, C. N. (2021). Antioxidant properties of anthocyanins and their mechanism of action in atherosclerosis. *Free Radic. Biol. Med.* 172, 152–166. doi: 10.1016/j.freeradbiomed.2021.05.040
- Garrison, E. P. (2015). *vcflib*. In: *GitHub repository* (GitHub). Available online at: <https://github.com/ekg/vcflib> (Accessed February 10, 2024).
- Garrison, E. P., and Marth, G. T. (2012). *Haplotype-based variant detection from short-read sequencing* (Ithaca, United States: Cornell University). Available at: <http://arxiv.org/abs/1207.3907> (Accessed February 09, 2024).
- He, F., Niu, M. X., Feng, C. H., Li, H. G., Su, Y., Su, W. L., et al. (2020). PeSTZ1 confers salt stress tolerance by scavenging the accumulation of ROS through regulating the expression of PeZAT12 and PeAPX2 in *Populus*. *Tree Physiol.* 40, 1292–1311. doi: 10.1093/treephys/tpaa050
- Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4, e1000117. doi: 10.1371/journal.pcbi.1000117
- Hugueney, P., Provenzano, S., Verriès, C., Ferrandino, A., Meudec, E., Batelli, G., et al. (2009). A novel cation-dependent O-methyltransferase involved in anthocyanin methylation in grapevine. *Plant Physiol.* 150, 2057–2070. doi: 10.1104/pp.109.140376
- Ino, I., and Yamaguchi, M. A. (1993). Acetyl-coenzyme A: Anthocyanidin 3-glucoside acetyltransferase from flowers of *Zinnia elegans*. *Phytochemistry* 33, 1415–1417. doi: 10.1016/0031-9422(93)85101-V
- IPCC (2021). "Summary for policymakers," in *Climate Change 2021. The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, vol. Vol. 2. Eds. V. Masson-Delmotte, P. Zhai and A. Pirani (Cambridge, United Kingdom: Cambridge University Press).
- Iqbal, M. A., Soren, K. R., Gangwar, P., Shanmugavadivel, P. S., Aravind, K., Singla, D., et al. (2017). Discovery of putative herbicide resistance genes and its regulatory network in chickpea using transcriptome sequencing. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.00958
- 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
- Jordanovska, S., Jovovic, Z., and Andjelkovic, V. (2020). "Potential of wild species in the scenario of climate change," in *Rediscovery of Genetic and Genomic Resources for Future Food Security*. Eds. R. K. Salgotra and S. M. Zargar (Springer Nature, Singapore), 263–302. doi: 10.1007/978-981-15-0156-2_10
- Ju, Y., Yang, B., He, S., Tu, T., Min, Z., Fang, Y., et al. (2019). Anthocyanin accumulation and biosynthesis are modulated by regulated deficit irrigation in Cabernet Sauvignon (*Vitis vinifera* L.) grapes and wines. *Plant Physiol. Biochem.* 135, 469–479. doi: 10.1016/j.plaphy.2018.11.013
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Koyama, R., Yoshimoto, A., Ishibashi, M., Itoh, H., and Uno, Y. (2021). Enzymatic activities and gene transcript levels associated with the augmentation of antioxidant constituents during drought stress in lettuce. *Horticulturae* 7, 444. doi: 10.3390/horticulturae7110444

- Langfelder, P., and Horvath, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinf.* 9, 559. doi: 10.1186/1471-2105-9-559
- Lee, S., Choi, E., Kim, T., Hwang, J., and Lee, J.-H. (2022). AtHAD1, a haloacid dehalogenase-like phosphatase, is involved in repressing the ABA response. *Biochem. Biophys. Res. Commun.* 587, 119–125. doi: 10.1016/j.bbrc.2021.11.095
- Li, M., Liu, J., Zhou, Y., Zhou, S., Zhang, S., Tong, H., et al. (2020). Transcriptome and metabolome profiling unveiled mechanisms of tea (*Camellia sinensis*) quality improvement by moderate drought on pre-harvest shoots. *Phytochemistry* 180, 112515. doi: 10.1016/j.phytochem.2020.112515
- Liao, Y., Smyth, G. K., and Shi, W. (2013). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656
- Llorach, R., Martínez-Sánchez, A., Tomás-Barberán, F. A., Gil, M. I., and Ferreres, F. (2008). Characterisation of polyphenols and antioxidant properties of five lettuce varieties and escarole. *Food Chem.* 108, 1028–1038. doi: 10.1016/j.foodchem.2007.11.032
- Medina-Lozano, I., Arnedo, M. S., Grimplet, J., and Diaz, A. (2023). Selection of novel reference genes by RNA-seq and their evaluation for normalising real-time qPCR expression data of anthocyanin-related genes in lettuce and wild relatives. *Int. J. Mol. Sci.* 24, 3052. doi: 10.3390/ijms24033052
- Medina-Lozano, I., Bertolin, J. R., and Diaz, A. (2021). Nutritional value of commercial and traditional lettuce (*Lactuca sativa* L.) and wild relatives: vitamin C and anthocyanin content. *Food Chem.* 359, 129864. doi: 10.1016/j.foodchem.2021.129864
- Medina-Lozano, I., Bertolin, J. R., and Diaz, A. (2024). Impact of drought stress on vitamin C and anthocyanin content in cultivated lettuces (*Lactuca sativa* L.) and wild relatives (*Lactuca* spp.). *Front. Plant Sci.* 15. doi: 10.3389/fpls.2024.1369658
- Moreno-Escamilla, J. O., Jiménez-Hernández, F. E., Alvarez-Parrilla, E., de la Rosa, L. A., Martínez-Ruiz, N. D. R., González-Fernández, R., et al. (2020). Effect of Elicitation on Polyphenol and Carotenoid Metabolism in Butterhead Lettuce (*Lactuca sativa* var. capitata). *ACS Omega* 5, 11535–11546. doi: 10.1021/acsomega.0c00680
- Mou, B. (2005). Genetic variation of beta-carotene and lutein contents in lettuce. *J. Am. Soc. Horticult. Sci.* 130, 870–876. doi: 10.21273/jashs.130.6.870
- Muñoz-Espinoza, C., Di Genova, A., Sánchez, A., Correa, J., Espinoza, A., Meneses, C., et al. (2020). Identification of SNPs and InDels associated with berry size in table grapes integrating genetic and transcriptomic approaches. *BMC Plant Biol.* 20, 1–21. doi: 10.1186/s12870-020-02564-4
- Na, J.-K., and Metzger, J. D. (2020). A putative tomato inositol polyphosphate 5-phosphatase, Le5PT1, is involved in plant growth and abiotic stress responses. *3 Biotech.* 10, 28. doi: 10.1007/s13205-019-2023-y
- Naing, A. H., and Kim, C. K. (2021). Abiotic stress-induced anthocyanins in plants: Their role in tolerance to abiotic stresses. *Physiol. Plant* 172, 1711–1723. doi: 10.1111/ppl.13373
- Paim, B. T., Crizel, R. L., Tatiane, S. J., Rodrigues, V. R., Rombaldi, C. V., and Galli, V. (2020). Mild drought stress has potential to improve lettuce yield and quality. *Sci. Hortic. (Amsterdam)* 272, 109578. doi: 10.1016/j.scienta.2020.109578
- Palusa, S. G., Golovkin, M., Shin, S. B., Richardson, D. N., and Reddy, A. S. N. (2007). Organ-specific, developmental, hormonal and stress regulation of expression of putative pectate lyase genes in *Arabidopsis*. *New Phytol.* 174, 537–550. doi: 10.1111/j.1469-8137.2007.02033.x
- Park, S., Shi, A., and Mou, B. (2020). Genome-wide identification and expression analysis of the CBF/DREB1 gene family in lettuce. *Sci. Rep.* 10, 5733. doi: 10.1038/s41598-020-62458-1
- Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29, e45. doi: 10.1093/nar/29.9.e45
- PGR (Plant Genetic Resources) Lettuce. *The lettuce gene pool*. Available online at: <https://www.pgrportal.nl/en/lettuce-genetic-resources-portal.htm> (Accessed January 27, 2024).
- Qu, J., Kang, S. G., Hah, C., and Jang, J. C. (2016). Molecular and cellular characterization of GA-Stimulated Transcripts GASA4 and GASA6 in *Arabidopsis thaliana*. *Plant Sci.* 246, 1–10. doi: 10.1016/j.plantsci.2016.01.009
- Quezada-Martínez, D., Addo Nyarko, C. P., Schiessl, S. V., and Mason, A. S. (2021). Using wild relatives and related species to build climate resilience in *Brassica* crops. *Theor. Appl. Genet.* 134, 1711–1728. doi: 10.1007/s00122-021-03793-3
- Ramírez, V., López, A., Mauch-Mani, B., Gil, M. J., and Vera, P. (2013). An extracellular subtilase switch for immune priming in *Arabidopsis*. *PLoS Pathog.* 9, e1003445. doi: 10.1371/journal.ppat.1003445
- Reddy, A. R., Chaitanya, K. V., and Vivekanandan, M. (2004). Drought-induced responses of photosynthesis and antioxidant metabolism in higher plants. *J. Plant Physiol.* 161, 1189–1202. doi: 10.1016/j.jplph.2004.01.013
- Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikat, S., Song, C., et al. (2017). Genome assembly with *in vitro* proximity ligation data and whole-genome triplication in lettuce. *Nat. Commun.* 8, 14953. doi: 10.1038/ncomms14953
- Ritossa, F. (1962). A new puffing pattern induced by temperature shock and DNP in drosophila. *Experientia* 18, 571–573. doi: 10.1007/BF02172188
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Rugienius, R., Bendokas, V., Siksnianas, T., Stanys, V., Sasnauskas, A., and Kazanavičiute, V. (2021). Characteristics of *Fragaria vesca* Yield Parameters and Anthocyanin Accumulation under Water Deficit Stress. *Plants* 10, 557. doi: 10.3390/plants10030557
- Shao, H. B., Chu, L. Y., Jaleel, C. A., Manivannan, P., Panneerselvam, R., and Shao, M. A. (2009). Understanding water deficit stress-induced changes in the basic metabolism of higher plants-biotechnologically and sustainably improving agriculture and the ecoenvironment in arid regions of the globe. *Crit. Rev. Biotechnol.* 29, 131–151. doi: 10.1080/07388550902869792
- Shim, J. S., Oh, N., Chung, P. J., Kim, Y. S., Choi, Y. D., and Kim, J. K. (2018). Overexpression of OsNAC14 improves drought tolerance in rice. *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00310
- Su, W., Tao, R., Liu, W., Yu, C., Yue, Z., He, S., et al. (2020). Characterization of four polymorphic genes controlling red leaf colour in lettuce that have undergone disruptive selection since domestication. *Plant Biotechnol. J.* 18, 479–490. doi: 10.1111/pbi.13213
- Tahmasebi, A., Ashrafi-Dehkordi, E., Shahriari, A. G., Mazloomi, S. M., and Ebrahimie, E. (2019). Integrative meta-analysis of transcriptomic responses to abiotic stress in cotton. *Prog. Biophys. Mol. Biol.* 146, 112–122. doi: 10.1016/j.pbmolbio.2019.02.005
- Tange, O. (2011). *GNU Parallel: The Command-Line Power Tool* Vol. 36 (Frederiksberg, Denmark: Login USENIX Mag), 42–47. Available at: <https://www.gnu.org/software/parallel/>. (Accessed February 09, 2024)
- The Galaxy Community (2022). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* 50, W345–W351. doi: 10.1093/nar/gkac247
- Tian, Y., Li, Q., Rao, S., Wang, A., Zhang, H., Wang, L., et al. (2021). Metabolic profiling and gene expression analysis provides insights into flavonoid and anthocyanin metabolism in poplar. *Tree Physiol.* 41, 1046–1064. doi: 10.1093/treephys/tpaa152
- Tsitsekian, D., Daras, G., Templalexis, D., Avgeri, F., Lotos, L., Orfanidou, C. G., et al. (2023). A subset of highly responsive transcription factors upon tomato infection by pepino mosaic virus. *Plant Biol.* 25, 529–540. doi: 10.1111/plb.13515
- Tsormpatsidis, E., Henbest, R. G. C., Davis, F. J., Battey, N. H., Hadley, P., and Wagstaffe, A. (2008). UV irradiance as a major influence on growth, development and secondary products of commercial importance in Lollo Rosso lettuce “Revolution” grown under polyethylene films. *Environ. Exp. Bot.* 63, 232–239. doi: 10.1016/j.envexpbot.2007.12.002
- Ul Haq, S., Khan, A., Ali, M., Khattak, A. M., Gai, W.-X., Zhang, H.-X., et al. (2019). Heat shock proteins: dynamic biomolecules to counter plant biotic and abiotic stresses. *Int. J. Mol. Sci.* 20, 5321. doi: 10.3390/ijms20215321
- Wada, K. C., Inagaki, N., Sakai, H., Yamashita, H., Nakai, Y., Fujimoto, Z., et al. (2022). Genetic effects of Red Lettuce Leaf genes on red coloration in leaf lettuce under artificial lighting conditions. *Plant Environ. Interact.* 3, 179–192. doi: 10.1002/pei3.10089
- Wang, Y., Chen, R., Hao, Y., Liu, H., Song, S., and Sun, G. (2017). Transcriptome analysis reveals differentially expressed genes (DEGs) related to lettuce (*Lactuca sativa*) treated by TiO₂/ZnO nanoparticles. *Plant Growth Regul.* 83, 13–25. doi: 10.1007/s10725-017-0280-5
- 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
- Wang, K., Guo, Q., Froehlich, J. E., Hersh, H. L., Zienkiewicz, A., Howe, G. A., et al. (2018). Two abscisic acid-responsive plastid lipase genes involved in jasmonic acid biosynthesis in *Arabidopsis thaliana*. *Plant Cell* 30, 1006–1022. doi: 10.1105/tpc.18.00250
- Wei, Z., Hu, K., Zhao, D. L., Tang, J., Huang, Z. Q., Jin, P., et al. (2020). MYB44 competitively inhibits the formation of the MYB340-bHLH2-NAC56 complex to regulate anthocyanin biosynthesis in purple-fleshed sweet potato. *BMC Plant Biol.* 20, 258. doi: 10.1186/s12870-020-02451-y
- Wickham, H. (2009). *ggplot2 - Elegant Graphics for Data Analysis* (New York, NY, USA: Springer). doi: 10.1007/978-3-319-24277-4
- Wu, N., Li, W. J., Chen, C., Zhao, Y. P., and Hou, Y. X. (2022). Analysis of the PRA1 genes in cotton identifies the role of GhPRA1.B1-1A in *Verticillium dahliae* resistance. *Genes (Basel)* 13, 765. doi: 10.3390/genes13050765
- Xiong, T., Zhang, S., Kang, Z., Zhang, T., and Li, S. (2021). Dose-Dependent Physiological and Transcriptomic Responses of Lettuce (*Lactuca sativa* L.) to Copper Oxide Nanoparticles-Insights into the Phytotoxicity Mechanisms. *Int. J. Mol. Sci.* 22, 3688. doi: 10.3390/ijms22073688
- Yan, H., Pei, X., Zhang, H., Li, X., Zhang, X., Zhao, M., et al. (2021). MYB-mediated regulation of anthocyanin biosynthesis. *Int. J. Mol. Sci.* 22, 3103. doi: 10.3390/ijms22063103
- Yazar, M., and Özbek, P. (2021). *In silico* tools and approaches for the prediction of functional and structural effects of single-nucleotide polymorphisms on proteins: an expert review. *Omi. A J. Integr. Biol.* 25, 23–37. doi: 10.1089/omi.2020.0141
- Zeljковиć, S. C., Štefelová, N., Hron, K., Doležalová, I., and Tarkowski, P. (2023). Preharvest abiotic stress affects the nutritional value of lettuce. *Agronomy* 13, 398. doi: 10.3390/agronomy13020398
- Zhang, F., Rosental, L., Ji, B., Brotman, Y., and Dai, M. (2024). Metabolite-mediated adaptation of crops to drought and the acquisition of tolerance. *Plant J.* 118, 626–644. doi: 10.1111/tjp.16634

Zhang, H., Wan, Z., Liu, J., Hu, X., Ren, L., Feng, S., et al. (2022). DsCER26 affects the leaf dehydration tolerance of rice by altering cuticular wax alkane production without affecting the grain fatty acid content. *ACS Agric. Sci. Technol.* 2, 813–822. doi: 10.1021/acscagritech.2c00141

Zhang, Y., Xu, S., Cheng, Y., Peng, Z., and Han, J. (2018). Transcriptome profiling of anthocyanin-related genes reveals effects of light intensity on anthocyanin biosynthesis in red leaf lettuce. *PeerJ* 13, e4607. doi: 10.7717/peerj.4607

Zhang, Y. Z., Xu, S. Z., Cheng, Y. W., Ya, H. Y., and Han, J. M. (2016). Transcriptome analysis and anthocyanin-related genes in red leaf lettuce. *Genet. Mol. Res.* 15, gmr.15017023. doi: 10.4238/gmr.15017023

Zhang, B., Zhu, Z. Z., Qu, D., Wang, B. C., Hao, N. N., Yang, Y. Z., et al. (2021). MdBBX21, a B-box protein, positively regulates light-induced anthocyanin accumulation in apple peel. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.774446

Zhou, H., Yu, L., Liu, S., Zhu, A., Yang, Y., Chen, C., et al. (2023). Transcriptome comparison analyses in UV-B induced AsA accumulation of *Lactuca sativa* L. *BMC Genomics* 24, 61. doi: 10.1186/s12864-023-09133-7

Zhou, T., Yue, C., Huang, J., Cui, J., Liu, Y., Wang, W., et al. (2021). Genome-wide identification of the amino acid permease genes and molecular characterization of their transcriptional responses to various nutrient stresses in allotetraploid rapeseed. *BMC Plant Biol.* 21, 151. doi: 10.1186/s12870-021-03043-0

Zhu, H., Xia, R., Zhao, B., An, Y., Dardick, C. D., Callahan, A. M., et al. (2012). Unique expression, processing regulation, and regulatory network of peach (*Prunus persica*) miRNAs. *BMC Plant Biol.* 12, 149. doi: 10.1186/1471-2229-12-149



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Raul Pirona,
National Research Council (CNR), Italy
Enéas Ricardo Konzen,
Federal University of Rio Grande do Sul, Brazil

*CORRESPONDENCE

Fernanda Carla Ferreira de Pontes
✉ fernandacfpones@alu.ufc.br

RECEIVED 01 June 2024

ACCEPTED 31 December 2024

PUBLISHED 23 January 2025

CITATION

de Pontes FCF, Machado IP, Silveira MVdS,
Lobo ALA, Sabadin F, Fritsche-Neto R and
DoVale JC (2025) Combining genotyping
approaches improves resolution for
association mapping: a case study in tropical
maize under water stress conditions.
Front. Plant Sci. 15:1442008.
doi: 10.3389/fpls.2024.1442008

COPYRIGHT

© 2025 de Pontes, Machado, Silveira, Lobo,
Sabadin, Fritsche-Neto and DoVale. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Combining genotyping approaches improves resolution for association mapping: a case study in tropical maize under water stress conditions

Fernanda Carla Ferreira de Pontes^{1*}, Ingrid Pinheiro Machado¹,
Maria Valnice de Souza Silveira¹, Antônio Lucas Aguiar Lobo¹,
Felipe Sabadin², Roberto Fritsche-Neto³
and Júlio César DoVale¹

¹Postgraduate Program of Plant Science, Federal University of Ceará, Fortaleza, Ceará, Brazil, ²College of Agriculture and Applied Sciences, Utah State University, Logan, UT, United States, ³Louisiana State University Agricultural Center, Baton Rouge, LA, United States

Genome-wide Association Studies (GWAS) identify genome variations related to specific phenotypes using Single Nucleotide Polymorphism (SNP) markers. Genotyping platforms like SNP-Array or sequencing-based techniques (GBS) can genotype samples with many SNPs. These approaches may bias tropical maize analyses due to reliance on the temperate line B73 as the reference genome. An alternative is a simulated genome called “Mock,” adapted to the population using bioinformatics. Recent studies show SNP-Array, GBS, and Mock yield similar results for population structure, heterotic groups definition, tester selection, and genomic hybrid prediction. However, no studies have examined the results generated by these different genotyping approaches for GWAS. This study aims to test the equivalence among the three genotyping scenarios in identifying significant effect genes in GWAS. To achieve this, maize was used as the model species, where SNP-Array genotyped 360 inbred lines from a public panel via the Affymetrix platform and GBS. The GBS data were used to perform SNP calling using the temperate inbred line B73 as the reference genome (GBS-B73) and a simulated genome “Mock” obtained *in-silico* (GBS-Mock). The study encompassed four above-ground traits with plants grown under two levels of water supply: well-watered (WW) and water-stressed (WS). In total, 46, 34, and 31 SNP were identified in the SNP-Array, GBS-B73, and GBS-Mock scenarios, respectively, across the two water levels, associated with the evaluated traits following the comparative analysis of each genotyping method individually. Overall, the identified candidate genes varied along the various scenarios but had the same functionality. Regarding SNP-Array and GBS-B73, genes with functional similarity were identified even without coincidence in the physical position of the SNPs. These genes and regions are involved in various processes and responses with applications in plant breeding. In terms of accuracy, the combination of genotyping scenarios compared to those isolated is feasible and recommended, as it increased all traits under both water conditions. In this sense,

it is worth highlighting the combination of GBS-B73 and GBS-Mock scenarios, not only due to the increase in the resolution of GWAS results but also the reduction of costs associated with genotyping and the possibility of conducting genomic breeding methods.

KEYWORDS

SNP-array, genotyping by sequencing, simulated genome, GWAS, genotyping platforms, candidate genes, genes with similar functions

1 Introduction

Water is the most abundant and often limiting of all plant resources needed to grow and function (Taiz et al., 2015). Water availability is considered one of the most influential factors in agricultural productivity, controlling species distribution in different climatic zones on Earth (Turner and Jones, 1980). In the tropical zone, characterized by relatively high temperatures and low rainfall compared to other zones, plants thriving in these environments are often more exposed to prolonged periods of water scarcity, especially in arid and semi-arid regions. According to climate change projections, this scenario will likely continue or worsen over the years, with potentially more drastic effects on plants (Raza et al., 2019; IPCC, 2023).

Stress can be considered a significant deviation from optimal life conditions (Larcher, 2003), inducing changes and responses as the plant fails to complete its physiological processes for growth and production. The lack of adequate water supply causes greater expansion of the root system into deeper and moister zones of the soil profile, reduction in the development of cells in the aerial tissues, resulting in decreased growth and stomatal closure to reduce transpiration rate and, consequently, photosynthetic activity (Frensch and Hsiao, 1994; Hsiao, 1973). Control measures are complex and difficult for humans to manage, and the search for genotypes that will perform better and economically viable yields in water-limited environments has been increasingly important for genetic improvement.

Conventional breeding for water deficit conditions is still time-consuming, laborious, and costly, as experimental conditions must be carefully managed. However, in recent years, with advances in molecular biology, the development of high-throughput genotyping technologies, and progress in platform development, new opportunities have emerged to enhance this process. This is partly due to cost reduction, which has consequently driven advances in genomic sequencing; another factor is the versatility of SNP (Single Nucleotide Polymorphism) markers, most commonly used in this process (Ingvarsson and Street, 2011). SNPs are abundant markers in crop genomes and are ideal for genetic discovery research and molecular improvement (Rasheed et al., 2017). According to the same authors, genotyping platforms involving Next Generation Sequencing (NGS) and SNP-Array technologies are suitable for

genotyping hundreds to thousands of samples with many SNP markers in a single assay much more quickly, revolutionizing the study of genomics and molecular biology.

Genotyping techniques by sequencing or GBS (Genotyping by Sequencing) are simple and highly multiplexed systems used for constructing libraries intended for next-generation sequencing. SNP-Array is a technique that uses microarrays designed to pre-select previously identified genetic markers characterized by wide polymorphism. These markers are then incorporated into a specific platform. GBS-scored SNP platforms provide many markers, although they have high rates of missing data. On the other hand, Array-scored SNP platforms are of high quality but have relatively high costs (Elbasyoni et al., 2018) and possible ascertainment bias if the genetic material used for array development is not related to the tested germplasm (Heslot et al., 2013).

Arrays are well-designed and established in the market to assist studies and breeding programs of major commodity crops (Ganal et al., 2011). For minor crops, arrays are still rarely available, and researchers often rely on information from other crops that is already accessible. However, due to the high cost associated with array development, these platforms are preferably employed when it is possible to use a “universal” approach that applies to various germplasms. However, this can be challenging if researchers attempt to identify rare SNPs across various germplasms; a universal design can become large and expensive, resulting in many monomorphic loci for non-target germplasm groups (Thomson, 2014).

The advancement of model genome knowledge and the advent of next-generation sequencing techniques open up the possibility of a great leap in understanding the genome of relatively lesser-known species. The GBS pipelines are based on a reference genome or assembly of a new genome, applied to model organisms and species lacking pre-existing genomic information (Davey et al., 2011; Poland et al., 2012). In cases where a reference genome is not yet available, a simulated genome can be employed for SNP discovery, which can serve as a valid alternative (Melo et al., 2016). The same authors developed a bioinformatics pipeline to construct a simulated genome called “Mock,” adapted to the population and built from GBS data. This genome is already being used in genomic studies and indicated that the Mock produces similar results when it comes to organizing populations, identifying heterotic groups,

selecting testers, and predicting genomic characteristics of hybrids compared to standard approaches (SNP-Array and GBS) (MaChado et al., 2023; Sabadin et al., 2022). This suggests that simulated genomes can be a good alternative, especially for species without the reference genome. However, no studies have been identified on the results generated by these different genotyping approaches in Genome-Wide Association Studies (GWAS).

Other studies have compared datasets from different high-throughput genotyping technologies in GWAS. Darrier et al. (2019), using standard platforms, GBS and SNP-Array, demonstrated efficiency in characterizing genetic diversity in barley, although accessing different regions of the genome. Despite capturing different areas, there was a positive correlation between the genetic distance matrices of both approaches, validating the use of either one for the characterization. These authors emphasized that the choice between GBS and SNP-Array genotyping platforms should be based on various factors, including the nature of the research and group preferences. For example, GBS may be preferable for studies requiring broader genomic coverage due to its ability to sequence a large number of genetic markers.

Conversely, SNP-Array may be more appropriate for analyses focused on specific genome regions. Group preferences, previous experience, and practical considerations such as cost and resource availability influence platform choice. In a study with inbred maize lines, Negro et al. (2019) concluded that GBS and SNP-Array were complementary for detecting QTL marking different haplotypes in association studies. Assuming they are complementary, combining these platforms seeks to determine if it will result in greater data accuracy.

To date, no study comparing GBS, SNP-Array, and simulated genome for GWAS has been published yet. The application of studies of this nature is crucial because they provide evidence that the information obtained from various genotyping approaches may be complementary during the genotyping process, thus demonstrating an efficient alternative for identifying polymorphisms. This, in turn, should offer better support to breeding programs that consistently grapple with identifying more efficient and tolerant genotypes against various abiotic and biotic factors. Another relevant point is that, even with advances in whole-genome sequencing and the complete publication of the maize genome, approaches such as SNP-Array and GBS remain important due to their lower cost and efficiency in genotyping large populations. These techniques generate more manageable data, requiring less computational infrastructure, and provide sufficient resolution to address many biological questions. Their effectiveness in GWAS studies and identifying loci in crops like maize is well demonstrated. For breeding programs or projects with limited resources, they represent agile and viable alternatives, balancing cost, accessibility, and quality.

In this context, the objectives of this study were: i) to verify if there is a difference in the identification of genes with significant effects among genotyping platforms, SNP-Array, GBS, and simulated genome ("Mock") in GWAS; ii) once differences are confirmed, to determine if the identified genomic regions are complementary and if they provide better accuracy.

2 Materials and methods

To enhance the comprehensibility of the analyses conducted in this study, we present a workflow in which the experimental and data analysis components are summarized in Figure 1. The subsequent sections provide detailed explanations.

2.1 Genetic material and experimental trials

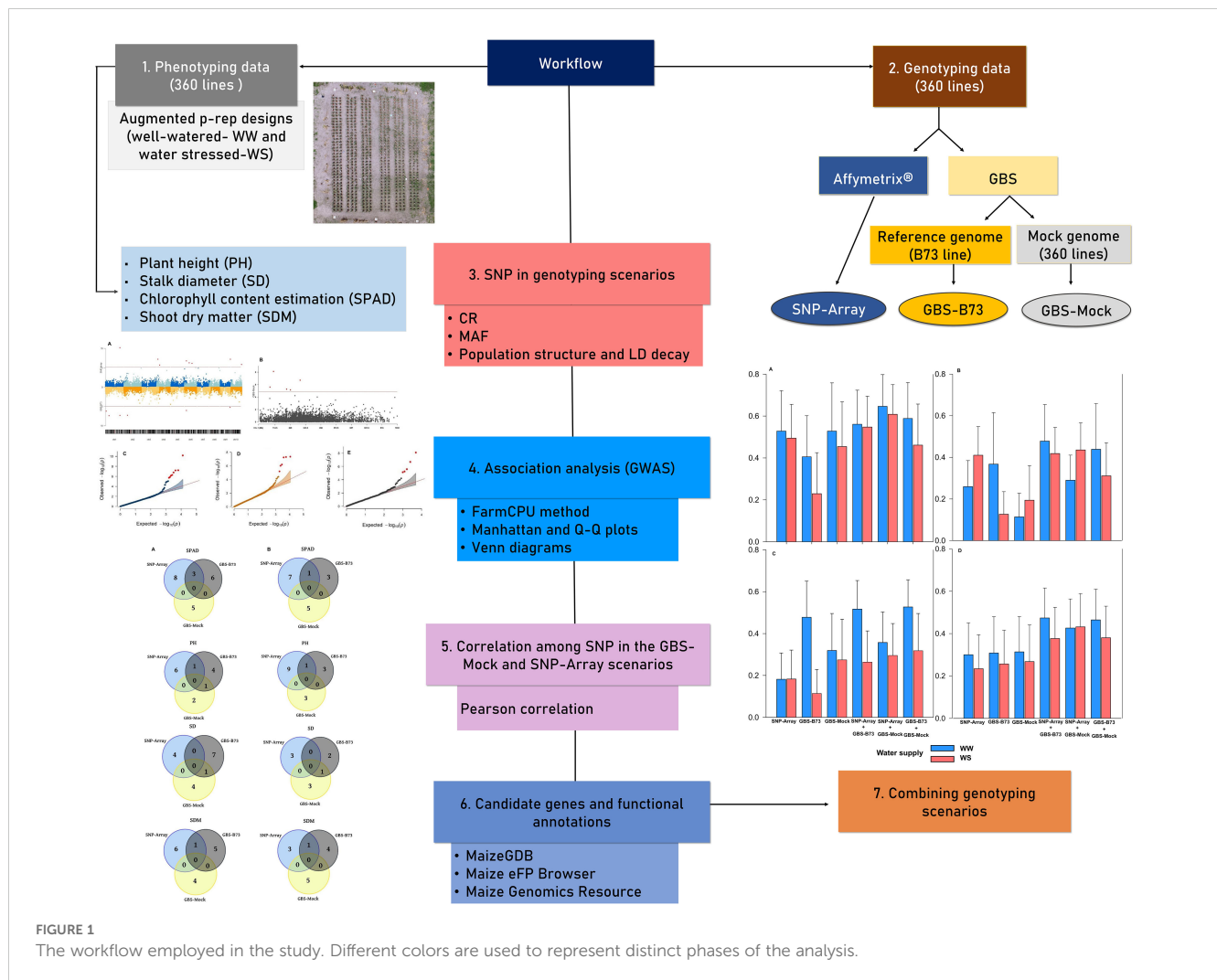
This study used maize as the model species in a public diversity panel of 360 tropical inbred lines (Yassue et al., 2021). The genomic and phenotypic data of this panel can be found on the Mendeley platform (<https://data.mendeley.com/datasets/6pb9prbb/1>). The data to be explored were obtained from eight experiments conducted in 2020 and 2021, as detailed below. This study involves contrasting water supply conditions, well-watered (WW) and water-stressed (WS), so a pilot experiment was conducted before the main experiments. A water retention curve was established through regression to obtain field capacity and determine the amount of water to be provided via irrigation (De Souza Silveira et al., 2024). This pilot experiment involved five randomly selected lines from the panel and five levels of water supply: 100% of water applied (WA), 80% of WA, 70% of WA, 50% of WA, and 40% of WA. As a result, the WW and WS points were determined, with the 80% WA and 40% WA treatments representing these conditions, respectively.

The main experiments were conducted at experimental fields of the Department of Agriculture at UFC, Campus do Pici, Fortaleza-CE, located at 3°44'24.27" S latitude and 38°34'29.93" W longitude. The main experiments were conducted under WW and WS in augmented partially repeated block design (augmented p-rep designs), with two temporally spaced replicates (Williams et al., 2011). Five common treatments (checks) were used, randomly selected from within the panel and distributed in each block within the WW and WS conditions (Supplementary Figure S1).

These experiments were always conducted in the second semester of each year, following the rainy season in the region, a period that resembles the climate of the semi-arid zone. The sowings were carried out in plastic pots with a capacity of 2000 cm³, containing substrate (easily reproducible) in a ratio of 3:1 (sand: earthworm humus). The earthworm humus was chosen due to its easy obtainability and effectiveness in providing nutrients to the plants. The use of sand is justified by its easy acquisition, availability, and low cost.

Two seeds were sown per pot at an average 3-4 cm depth. Thinning was performed when the seedlings reached the V2 stage, leaving only one seedling per pot (plot). At this same phenological stage, a water deficit was also initiated, which continued until the V6 stage (harvest). Planting and topdressing fertilization were based on the chemical analysis of the substrate, taking into consideration the crop recommendations, to isolate nutritional stress during the experimental conduct.

As the experiment was conducted in an open field, irrigation control for each experiment was carried out manually and daily.



Thus, 15 random samples were used to calculate the daily average weight of the pots within each water supply level. Subsequently, the difference between the current and total weights obtained at each water supply level was calculated to replenish the water volume. It is worth noting that, for each vegetative stage, the average plant weight was obtained to subtract it along with the current weight, thus not affecting the volume of water to be replenished.

2.2 Phenotypic data

The phenotypic evaluation was conducted when most plants reached the V6 phenological stage. The traits considered in this study were: Plant height (PH) - measured from the soil to the insertion of the flag leaf, measured using a graduated ruler (cm); Stalk diameter (SD) - an average of two measurements above ground level at the second node of the stem obtained using a caliper (mm); Chlorophyll content estimation - using SPAD, measuring three leaves per plant to get the average. Subsequently, the plants were cut off at ground level, placed in paper bags, and put in a forced-air oven at 65°C for 72 hours to obtain: Shoot dry matter (SDM)- quantified using an electronic analytical balance (0,005 g).

2.3 Phenotypic analysis

The outliers of the phenotypic data for the traits described in section 2.2 were removed. Then, the remaining data were adjusted for normality using the *bestNormalize* package (Peterson, 2021), and the assumptions of normal distribution were checked via the Shapiro test and Q-Q plots. Subsequently, equations of mixed linear models were fitted to obtain the BLUP by REML for each trait studied under WW and WS conditions, using the *sommer* package (Covarrubias-Pazarán, 2016). These analyses were performed using the following model:

$$y = X_1t + X_2l + X_3n + Z_1b + Z_2g + Z_3i + \epsilon \quad (1)$$

where, y is the vector of phenotypic values of the inbred lines panel and checks; X_1 , X_2 , and X_3 are incidence matrices for t , l , and n fixed effects; Z_1 , Z_2 and Z_3 are incidence matrices for b , g e i random effects; t is the water supply fixed effect vector (WW and WS conditions); l is the replicate (season) fixed effect vector within water supply; n is the number of leaves used as a covariate to correct for differences in plant vigor and development; b is the block/water supply/season random effect vector, where $g \sim N(0, I\sigma_b^2)$; g s the genotype random effect vector, where $g \sim N(0, I\sigma_g^2)$; i is the random

effect vector of the genotype–water supply interaction, where $i \sim N(0, I\sigma_i^2)$; ϵ is the experimental error, where $\epsilon \sim N(0, R\sigma_\epsilon^2)$, obtained using a structured diagonal matrix to make it possible to estimate two residual variances, one for each water supply level (σ_{eWW}^2 and σ_{eWS}^2). The significance of fixed effects was assessed using the Wald test, and random effects using the likelihood ratio test.

The variance components were used to estimate the heritabilities (h^2) by the following estimator:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{s} + \frac{(\sigma_{eWW}^2 + \sigma_{eWS}^2)}{rs}} \quad (2)$$

where h^2 refers to the entry-mean heritability; σ_g^2 is the genotypic variance of the inbred lines panel, σ_{ge}^2 is the variance of the genotype–water supply interaction; σ_{eWW}^2 and σ_{eWS}^2 are the environmental variance components in WW and WS; s are levels of WW and WS; and r is the number of repetitions in each water supply level.

The reliability of selection for each line [$R^2(\hat{\alpha}_i)$] was obtained by the following expression (Gorjanc et al., 2015):

$$R^2(\hat{\alpha}_i) = 1 - \frac{Var(\alpha_i - \hat{\alpha}_i)}{Var(\alpha_i)} \quad (3)$$

where $Var(\alpha_i - \hat{\alpha}_i)$ is the variance of the prediction error (PEV) of line i and $Var(\alpha_i)$ is the genotypic variance of the trait.

The de-regressed BLUPs (dBLUPs) were obtained by calculating the ratio between the BLUPs of each inbred line in WW and WS and their respective average reliabilities. After these analyses, 313 lines remained out of the 360 in the panel. The dBLUPs of these lines in WW and WS were used in the GWAS analyses.

2.4 Genotypic data

The lines were genotyped using two SNP genotyping platforms: Affymetrix® Axiom Maize Genotyping Array with 18,413 SNP markers (SNP-Array) and genotyping-by-sequencing (GBS) process following the sequencing protocol established by Poland et al. (2012). In this method, genomic DNA was digested by two restriction enzymes, PstI and MseI, to reduce the genome complexity. Subsequently, specific adapters for sequencing on the Illumina NextSeq 500 platform (Illumina Inc., San Diego, CA, United States) were attached to the digested fragments.

The primary GBS data were employed for two purposes: firstly, to perform SNP calling using the temperate line B73 as the reference genome (RefGen v4). Secondly, to construct a simulated reference genome (mock genome) for SNP calling, following the pipeline proposed by Melo et al. (2016), considering all the lines in the panel (Mock).

Therefore, the SNP data were subjected to three GWAS approaches: 1) SNP-Array; 2) GBS with SNP calling based on the B73 reference genome (GBS-B73); 3) GBS using the simulated genome as the reference (GBS-Mock). The SNPs for the GBS dataset was identified from raw data using the TASSEL 5.0 GBSv2 pipeline (Glaubitz et al., 2014), considering both GBS-B73 and

GBS-Mock as reference genomes, employing the BWA aligner. The BWA aligner (Li and Durbin, 2009) was used to align the tags against the reference genome (GBS-B73 and GBS-Mock).

The SNP sets obtained in these scenarios were submitted to quality control parameters as call rate (CR) and Minor Allele Frequency (MAF) procedures, where markers with CR < 90% and MAF lower than 5%, and non-biallelic markers were removed from the datasets (Morosini et al., 2017). Imputation of missing data was performed using the Beagle 5.0 algorithm (Browning and Browning, 2008).

2.5 Population structure and LD decay

In order to minimize potential bias caused by population structure, a PCA was performed based on the additive genomic relationship matrix among the remaining 313 panel lines, following VanRaden (2008) using the *SNPRelate* package (Zheng et al., 2012). FarmCPU automatically incorporated the correction via PCA in the association analysis. Two principal components were used to correct the population structure effect, and the best fit for the model was determined based on Q-Q plots. The most likely number of groups within the panel was determined according to Yassue et al. (2021) as it involved the same diversity panel.

The Linkage Disequilibrium (LD) estimation between each pair of SNP within the chromosomes was calculated by the square of the allele frequency correlation (r^2) among all SNP within a distance less than 1 Mbp. The r^2 values were plotted against the base pair distance of the SNP pair to obtain the LD decay by chromosome. This procedure was performed with all SNP retained from the quality control procedures.

2.6 Association analysis

GWAS were performed for each trait under WW and WS conditions using the FarmCPU method (Liu et al., 2016a). The method stands out for its computational efficiency and ability to control false positives, demonstrating greater statistical power in situations where the trait is strongly associated with kinship (Liu et al., 2016a; Segura et al., 2012). The *FarmCPU.P.Threshold* function was employed to obtain the p -threshold, specific for each trait via a simulation process with 100 permutations. Subsequently, the cutoff point was obtained by the ratio between the p -threshold and the number of markers used. Subsequently, p -values (significance), MAF, and ASE (Average Effect of Allele Substitution) were obtained for each significantly associated SNP, designated hereafter as a potential candidate gene underlying the target trait. Furthermore, the coefficient of determination for each significant SNP (R_{SNP}^2) was obtained based on ASE and MAF using equations described in Da et al. (2014). Next, multiple linear regressions were established for each trait using the significant SNPs as predictor variables to quantify the markers' influence on that trait's expression (R_{TOT}^2). The Manhattan and Q-Q plots graphs were generated using the CMplot package (Yin, 2020), and the graphs showing the proportion of phenotypic variance explained by

the SNP were generated using the ggplot2 package (Wickham, 2011) in the R software. Venn diagrams based on the common gene functionality for the traits at each water supply level were created using LucidChart (lucidchart.com).

2.7 Correlation among markers of different scenarios

Given the stability and efficiency of SNP-Array technology in accurately genotyping numerous markers, we conducted Pearson correlation analysis (r) among significant markers with known functions identified in GWAS within the GBS-Mock scenario and markers present in the SNP-Array scenario for each trait under both WW and WS conditions. The analyses were performed using the R software base. This approach aimed to assess the concordance and potential overlap between markers identified through different genotyping methods and their associations with specific traits. By comparing these markers across scenarios, we sought to elucidate common genetic factors contributing to trait variation and explore the utility of integrating data from diverse genotyping platforms in genomic analyses related to crop improvement and adaptation to environmental stressors.

2.8 Gene annotation

A candidate gene association mapping was performed for traits with significant SNP. The physical positions of SNP for GBS-Mock were assigned using BLAST (Altschul et al., 1990) to align them with the maize genome assembly for comparison purposes. These positions were used to obtain 41 bp DNA fragments on a single chromosome (Supplementary Table S2). Subsequently, a BLAST was conducted exclusively for GBS-Mock on MaizeGDB¹ via blast, utilizing the B73 RefGen_v4 sequence database to locate the chromosome by inserting the DNA fragment. The MaizeGDB database and its functional information associated with each SNP based on B73 RefGen_v4 were utilized for all scenarios. After defining the region to be considered, potential candidate genes flanking each marker were identified. Candidate genes linked to each trait were determined through annotation within a sliding window of 50 kb around each significant SNP, following a conservative approach described by Yassue et al. (2021). All genes within a range of 50 kb downstream and 50 kb upstream were annotated. Subsequently, they were assessed and considered based on two criteria: proximity to the SNP and functional similarity as per databases available on the Maize eFP Browser (2023)² and Maize Genomics Resource (2023)³.

1 maizegdb.org
2 bar.utoronto.ca
3 maize.uga.edu

3 Results

3.1 Phenotypic analysis

In general, significant effects were detected for all sources of variation, except for the G x WA interaction, in the studied traits (Table 1). The variance components showed a similar pattern for all traits, with a predominance of genotypic variance over the residual variance of the interaction. Except for the PH trait, there was a higher residual variance for the well-watered environment than the low-water availability. The genotypic variance component ranged from 0.09 to 0.19, and the genotype x environment interaction approached zero for all traits, affecting the estimates of heritabilities and accuracy. Heritabilities ranged from moderate to high magnitude, ranging from 0.58 to 0.73. PH was the trait with the highest heritability (0.73) and the least influenced by the environment, showing the highest genotypic coefficient of variation (0.197). The adjusted means fall within the same range observed in other studies.

3.2 Genotypic scenarios: number and distribution of SNP

After the quality control, heterozygous markers were eliminated using the MAF and CR procedures, resulting in 12.704 SNP

TABLE 1 Wald test of fixed effects, likelihood-ratio test (LRT) of random effects, variance components, heritability, accuracy, and adjusted average for SPAD, plant height (PH), stalk diameter (SD), and shoot dry matter (SDM) of the inbred lines evaluated in WW (well-watered) and WS (water-stressed) conditions water supply.

Source of variation	SPAD	PH	SD	SDM
	Wald statistic			
Water supply (WA)	1098.27***	4353.03***	4549.80***	5393.32***
Replicates/WA	1076.83***	2542.48***	1842.71***	2462.96***
Likelihood-ratio test (LRT)				
Block/WA/Season	89.35***	40.76***	46.21***	212.23***
Genotypes (G)	102.59***	201.98***	100.60***	96.35***
G x WA	1.11 ^{NS}	1.08 ^{NS}	0.72 ^{NS}	2.43 ^{NS}
Variance components				
σ_g^2	0.192	0.197	0.119	0.099
$\sigma_{g \times e}^2$	0.015	0.007	0.007	0.012
σ_{eWW}^2	0.525	0.266	0.352	0.299
σ_{eWS}^2	0.511	0.279	0.275	0.240
Heritability and accuracy				
h^2	0.58	0.73	0.59	0.58
Adjusted means				
	31.13	9.44	7.56	2.32

***significant at the 0.001 probability level (by Wald test or LRT), respectively.
^{NS}non-significant.

markers for SNP-Array out of a total of 18,413, 11,153 out of 131,350 for GBS-B73, and 4,935 out of 46,926 for GBS-Mock, which were used in the association analyses (Table 2). Approximately 69% of the marker set remained in the SNP-Array, while 10.5% remained in the GBS-Mock and 8.5% in the GBS-B73 scenario. However, there was a balanced distribution of SNP across the chromosomes in the standard scenarios (SNP-Array and GBS-B73). In total, 11 common SNPs were found among the genotypic scenarios (Figure 2). Between SNP-Array and GBS-B73, 8 shared SNPs were observed: 3 for SPAD under well-watered conditions and 1 under water-stressed conditions; 1 SNP in each water supply condition for PH and SDM. In the GBS-B73 and GBS-Mock scenarios, 1 common SNP was identified for PH under well-watered conditions and 1 for SD in both water supply conditions.

3.3 GWAS analysis

Significant SNP were found on five of the ten maize chromosomes for the SNP-Array scenario and four for GBS-B73 for the SPAD trait under the WW condition (Figures 3A, B). The Q-Q plots showed data fitted to the model (Figures 3C–E). The significant marker/trait association threshold ranged from 4.99 to 12.85 (Supplementary Table S1).

A total of 46, 34, and 31 significant SNP were found for SNP-Array, GBS-B73, and GBS-Mock, respectively (Table 3; Supplementary Figures S2–S9). There were no SNP common to all three scenarios; however, at least one SNP was shared between two of them (Figure 2). SPAD had the highest number of significant SNP, totaling 34, followed by PH, SDM, 27, and SD, 23. The SNP

array presented more markers for SPAD and PH and GBS-B73 for SD, and there was an equivalence among the three scenarios for SDM. Overall, GBS-B73 and GBS-Mock showed some similarity in the quantity of markers.

3.4 Correlation among SNP in the GBS-Mock and SNP-Array scenarios

Our results revealed 20 significant markers in the GBS-Mock that positively correlated with the SNP-Array scenario to traits under different environmental conditions (Table 4). Pearson correlation coefficients (r) were observed, ranging from weak to strong. Specifically, for SDM in WW conditions, correlations ranged from 0.94 to 0.30. Similarly, SPAD values showed moderate to strong correlations with markers, ranging from 0.52 to 0.76 in WW conditions and from 0.40 to 0.76 in WS conditions. For PH, correlations were moderate, with values of 0.36 for WW and 0.51 and 0.53 for WS. Notably, SD exhibited correlations ranging from 0.35 to 0.87 in WW conditions and from 0.30 to 0.87 in WS conditions. Additionally, SDM showed moderate to strong correlations, ranging from 0.46 to 0.94 in WW conditions and 0.47 in WS conditions.

3.5 Candidate genes and functional annotations

Based on the physical location of significant SNP in the B73 reference genome for SNP-Array and GBS-B73 and the reference genome for GBS-Mock, genomic regions and candidate genes related to significant loci were identified (Supplementary Table S1). In some cases, the same genes and regions were identified for a given trait under both water supply conditions. For example, *Zm00001d042735* and *Zm00001d001852* in the GBS-Mock scenario for SPAD and SD, respectively; *Zm00001d017978* located on chromosome 5 in SNP-Array for PH. Similarly, identical genes and regions were found in different scenarios, for instance, *Zm00001d031759* located on chromosome 1 was detected in SNP-Array and GBS-B73 for SPAD in WW and WS. The same gene was also identified for different traits, such as *Zm00001d005090* for SD and SDM in GBS-B73.

The genomic regions and candidate genes with similar functions were grouped, considering each trait at the same water supply level across genotyping scenarios (Figure 2; Supplementary Table S3). For SNP-Array and GBS-B73, regions and genes with the same functionality on the same chromosome were observed, such as *Zm00001d031445* and *Zm00001d027626*, both on chromosome 1, which are correlated with ethylene biosynthesis for SDM in WW. Conversely, these platforms also identified genomic regions and candidate genes on different chromosomes but with coinciding functions. For example, *Zm00001d026477* on chromosome 10 and *Zm00001d027695* on chromosome 1 are responsible for responses to abiotic stress by reactive oxygen species (ROS), jasmonic acid (JA), and ethylene; *Zm00001d044194* on chromosome 3 and *Zm00001d018127* on chromosome 5 function in the regulation of

TABLE 2 Number of markers scored (raw data) and the final number of markers (clean data) total and per chromosome (Chr) after quality control for all genotyping scenarios used to assess inbred lines evaluated in WW (well-watered) and WS (water-stressed) conditions water supply.

	Genotyping scenarios ^a		
	SNP-Array	GBS-B73	GBS-Mock
Raw data	18,413	131,350	46,926
Clean data	12,704	11,153	4,935
Chrm 1	1,977 (15.6%)	1,651 (15.0%)	Unique chrn
Chrm 2	1,643 (12.9%)	1,411 (12.7%)	
Chrm 3	1,430 (11.3%)	1,269 (11.4%)	
Chrm 4	1,412 (11.1%)	1,177 (10.6%)	
Chrm 5	1,373 (10.8%)	1,336 (12.0%)	
Chrm 6	1,018 (8.0%)	859 (7.7%)	
Chrm 7	957 (7.5%)	831 (7.5%)	
Chrm 8	1,116 (8.8%)	964 (8.6%)	
Chrm 9	973 (7.7%)	855 (7.7%)	
Chrm 10	805 (6.3%)	780 (7.0%)	

^aSNP-Array, Affymetrix® Axiom Maize Genotyping array; GBS-B73, genotyping-by-sequence with SNP calling using B73 as reference genome; GBS-Mock, genotyping-by-sequence with SNP calling using the mock reference built with all parental lines.

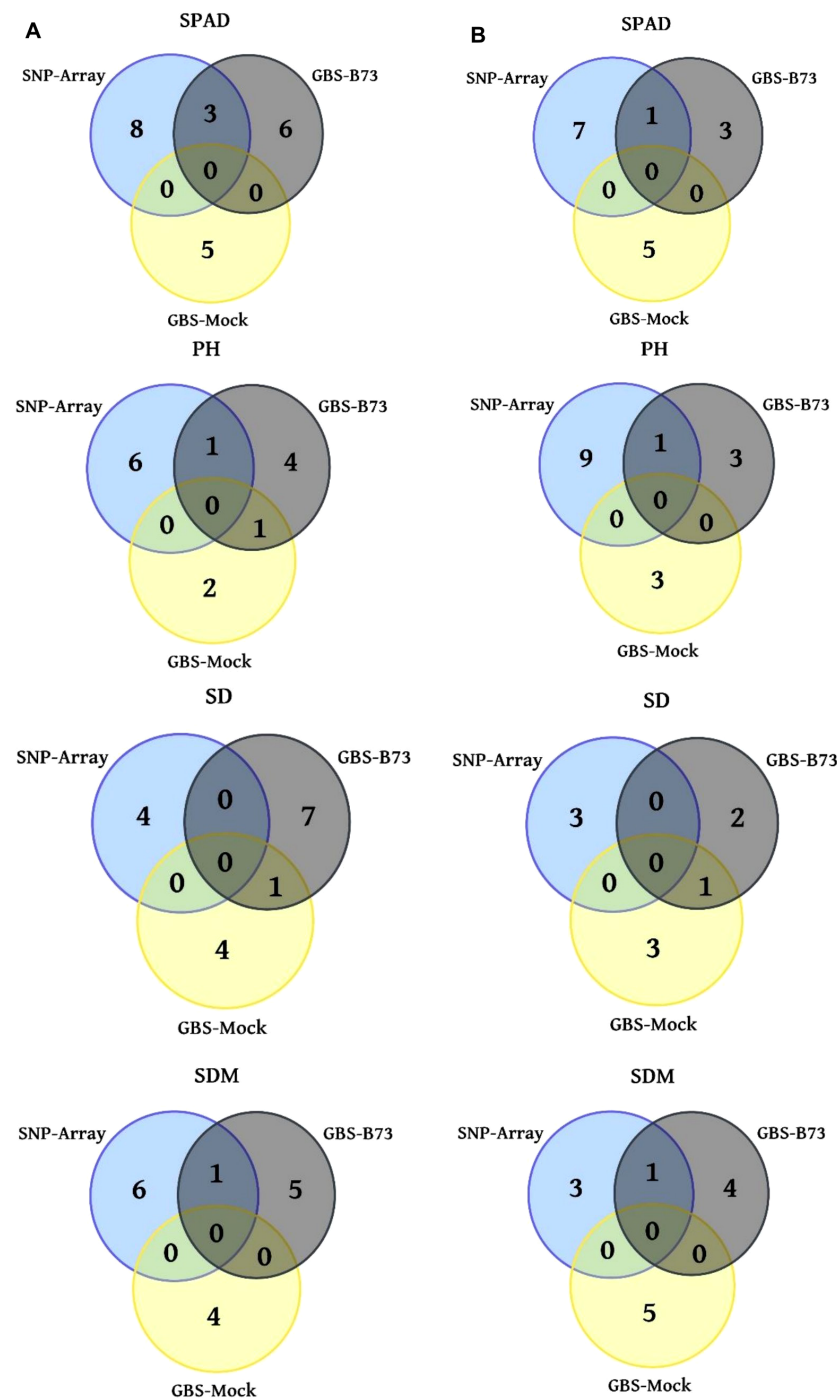


FIGURE 2

Venn diagrams with the number of significant trait SNPs in three genotyping scenarios. (A) WW (well-watered) water supply condition column; (B) WS (water-stressed) water supply condition column. SPAD, PH (plant height), SD (stalk diameter), and SDM (shoot dry matter).

the circadian cycle for SPAD under WW; *Zm00001d017978* on chromosome 5 and *Zm00001d008952* on chromosome 8 are involved in endoglucanase activity for PH in WW; and *Zm00001d053809* on chromosome 4 and *Zm00001d042481* on chromosome 3 for GBS-B73 are associated with ubiquitin proteins for PH in WS; *Zm00001d016786* on chromosome 5 and *Zm00001d005090* on chromosome 2 act in response to water stress through abscisic acid (ABA) for SDM in WS.

In scenarios involving GBS-B73 and GBS-Mock, genomic regions and candidate genes with similar functions were identified for *Zm00001d021708* on chromosome 7 and *Zm00001d012719* on the single chromosome, related to plant responses to ABA for PH in WW; *Zm00001d014899* on chromosome 5 and *Zm00001d001852* on the single chromosome, associated with the phytohormone gibberellin for SD in WW; *Zm00001d00509* on chromosome 2 and *Zm00001d053262* on the single chromosome, involved in ABA regulation for SD in WS.

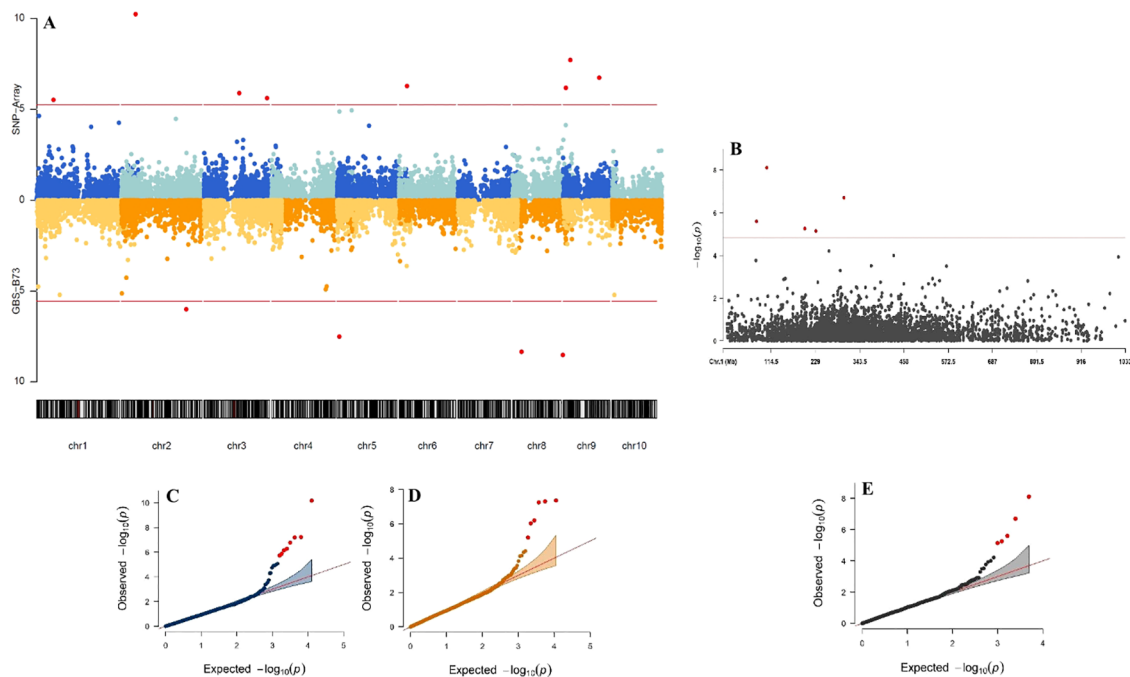


FIGURE 3

Manhattan plot and Quantile-Quantile (Q-Q) plots for Genome-Wide Association Study (GWAS) comparing genotyping platforms in tropical maize for SPAD trait in WW (water-stressed) conditions water supply. The Manhattan plot displays GWAS results based on three datasets: SNP-Array and GBS-B73 (A), and Mock (B). The x-axis represents the chromosomal positions, while the y-axis indicates the $-\log_{10}$ P-values, reflecting statistical significance. The horizontal lines denote the genome-wide suggestive significance threshold, with dots above these lines marking significant SNPs. The Q-Q plots illustrate the GWAS results for the same datasets: SNP-Array (C), GBS-B73 (D), and Mock (E). The x-axis corresponds to the $-\log_{10}$ expected P-values derived from the chi-square distribution, while the y-axis represents the $-\log_{10}$ observed P-values. Each dot represents an SNP, with the most significant SNP appearing as the top hit. The red diagonal line shows the expected distribution under the null hypothesis of no association.

3.6 Phenotypic variation explained by SNP in different genotyping scenarios

The proportions of phenotypic variance explained by significant SNP (R^2_{TOT}) for the analyzed traits under both water supply conditions, ideal (WW) and deficit (WS), were less explained in the isolated genotyping scenarios for the studied traits (Figure 4). Regarding the isolated scenarios, R^2_{TOT} in SNP-Array ranged from 0.18 for SD (WW in WS) to 0.53 for SPAD (WW), GBS-B73 ranged from 0.11 for SD (WS) to 0.48 for SD (WW), and GBS-Mock from 0.11 for PH (WW) to 0.53 for SPAD (WW). Overall, SNP-Array performed better independently for SPAD and PH, except for SD (WW), where GBS-B73 stood out, and SDM was almost the same among the scenarios. When combined, the value of R^2_{TOT} ranged from 0.26 in SNP-Array + GBS-B73 for SD (WS) to 0.65 in SNP-Array + GBS-Mock for SPAD (WW). The percentages obtained represent the phenotypic variance explained by combining multiple SNPs simultaneously.

The best scenario combination was SNP-Array + GBS-Mock for SPAD (WW) with an increase of 0.12 in accuracy compared to the best isolated scenario. For PH and SDM under WW condition, SNP-Array + GBS-B73 was superior, increasing accuracy by 0.07 and 0.16, respectively, compared to the best single scenario. For SD, combining GBS-B73 + GBS-Mock increased accuracy by 0.05. Regarding water availability, the ideal water supply condition

achieved better overall accuracy, except for PH in isolated SNP-Array and combined with GBS-Mock. In the WS condition, better accuracy was also observed for all traits when combining scenarios.

4 Discussion

Water is one of the most important factors limiting crop growth. Maize requires a large amount of water throughout all stages of development, from seed germination to the reproductive phase. In this context, the significant effect of water supply levels reveals contrasting conditions in WW and WS, indicating that the irrigation treatments used in the present study to generate contrasting environments were sufficient for all traits (Table 1). Moreover, the significance of genotypes suggests that the panel used in this study exhibits genetic variability. Previous studies have also reported genetic diversity for the same tropical maize germplasm panel (Yassue et al., 2021; De Souza Silveira et al., 2024). Genetic variability is a fundamental factor for any breeding program.

However, the interaction effect shows that the responses were not differentiated for the genotypes across environments; they exhibit similar phenotypic responses to environmental changes. Genotype $\times$ environment is important when estimating heritability because it influences a trait's genetic and environmental variation (Falconer and Mackay, 1996). The low effect of interaction also

TABLE 3 Number, average, and standard deviation (SD) of significant SNPs per trait in WW (well-watered) and WS (water-stressed) conditions water supply and genotyping scenario.

Water Supply	SPAD			PH			SD			SDM		
	SNP-Array	GBS-B73	GBS-Mock	SNP-Array	GBS-B73	GBS-Mock	SNP-Array	GBS-B73	GBS-Mock	SNP-Array	GBS-B73	GBS-Mock
WW	8	6	5	6	4	2	4	7	4	6	5	4
WS	7	3	5	9	3	3	3	2	3	3	4	5
Overall	15	9	10	15	7	5	7	9	7	9	9	9
Average	7.5	4.5	5	7.5	3.5	2.5	3.5	4.5	3.5	4.5	4.5	4.5
SD	0.5	1.5	0	1.5	0.5	0.5	0.5	2.5	0.5	1.5	0.5	0.5

TABLE 4 Pearson correlation among significant markers from the GBS-Mock scenario with known functions and markers from the SNP-Array scenario for traits in WW (well-watered) and WS (water-stressed) conditions water supply.

Trait	GBS-Mock		SNP-Array		r
	Marker	Position	Chrm	Position	
SPAD in WW	Zm00001d036175	76554426	6	66088433	0.71
			6	68443154	
	Zm00001d042735	201318938	6	76875108	0.76
			6	77806057	
	Zm00001d042755	230114762	1	99951335	0.41
SPAD in WS	Zm00001d024497	302570660	10	22670857	0.52
	Zm00001d008500	200328622	3	179265803	0.40
			4	214338748	
	Zm00001d042735	201318938	6	76875108	0.76
			6	77806057	
	Zm00001d006357	306796326	2	206059268	0.49
PH in WW	Zm00001d029023	537851163	5	199240894	0.41
	Zm00001d012719	305928723	9	1840217	0.36
PH in WS	Zm00001d034057	399525514	1	279982555	0.51
	Zm00001d018703	922415122	8	132436479	0.53
SD in WW	Zm00001d026300	205301556	1	62995980	0.38
	Zm00001d001852	583680097	7	119716831	0.35
	Zm00001d047956	831219654	10	144416974	0.87
SD in WS	Zm00001d001852	583680097	7	119716831	0.35
	Zm00001d053262	751364673	1	33596576	0.30
			1	156697884	
			2	2509863	
			2	176643239	
			2	226012848	
			3	200304796	
			4	30519074	

(Continued)

TABLE 4 Continued

Trait	GBS-Mock		SNP-Array		r
	Marker	Position	Chrm	Position	
			4	74777100	
			4	194400472	
			4	195155904	
			4	235373522	
			5	153728179	
			5	219174423	
			7	119716831	
			8	26683716	
			8	136183604	
SDM in WW	Zm00001d047956	831219654	10	144416974	0.87
	Zm00001d046354	320987424	9	84921243	0.94
	Zm00001d008954	511458983	10	94908425	0.46
SDM in WS	Zm00001d018001	440728659	5	211640184	0.47

maximizes the accuracy (Assareh et al., 2012); high accuracy estimates indicate good experimental precision. Heritability was higher for plant height, followed by stem diameter, consistent with results from Sabiel et al. (2013), who reported moderate heritabilities for plant height and stem diameter in maize under water stress.

4.1 SNP in genotyping scenarios

Platforms such as SNP-Array and GBS are well-suited for genotyping hundreds to thousands of samples, each containing numerous SNP markers, in a single assay and at a significantly faster pace (Rasheed et al., 2017). This study had a balanced SNP distribution across chromosomes in the SNP-Array and GBS-B73 genotyping scenarios, perhaps attributed to using the same reference genome (Table 2). The inbred line B73 has been utilized as the reference genome for maize sequencing (Schnable et al., 2009), and an example of a reference genome-based pipeline is TASSEL-GBS.

In the GBS-Mock scenario, a smaller number of SNP markers was observed. In cases where a reference genome is not yet available, a simulated genome can perform SNP discovery, serving as a valid alternative, especially for minor crops (MaChado et al., 2023; Sabadin et al., 2022). Regarding the smaller number of markers observed in GBS-B73 compared to SNP-Array, this may be related to the low genomic coverage of GBS resulting in missing SNP (Wang et al., 2020). However, this issue can be partially addressed by using software employed in imputation, as missing SNP are imputed to fill in the gaps in obtaining intermediate genotype information.

4.2 GWAS and candidate genes

GWAS has emerged as a crucial tool, allowing for a systematic approach to identifying associations between thousands of genomic loci and complex traits. In this study, overall, more SNP were identified in association with the trait under ideal water supply conditions than under water deficit conditions in all genotyping scenarios (Table 3). A similar result was found by De Souza Silveira et al (2024), who identified more SNPs associated with root traits of tropical maize under ideal water supply conditions than those subjected to water scarcity. Moreover, Yassue et al (2021); Yassue et al, 2023) found more SNP associated with tropical maize traits not evaluated under inoculation by growth-promoting bacteria, such as plant height, stem diameter, and aboveground dry mass. These authors also consider that growth-related traits, such as plant height, stem diameter, and dry mass, are complex and controlled by many genes with small individual effects.

The genes found in the study have small effects (ASE), revealing the polygenic nature of the traits and controlling a relatively small portion of the genotypic variation (Supplementary Table S1). Complex traits in plants, such as height, diameter, and tolerance to environmental stresses, often have a multifactorial genetic basis involving the interaction of various genes and environmental factors. Thus, knowledge of the genomic regions associated with the traits of interest will provide insight into this genetic basis. Additionally, the study also detected a common marker associated with more than one trait at different water supply levels, indicating a possible pleiotropic effect. Bouchet et al. (2017) reported pleiotropy among phenology-related traits, such as plant height and leaf number, and Zhang et al. (2022) for maize productivity traits. Pleiotropic effects in GWAS studies can increase the complexity of

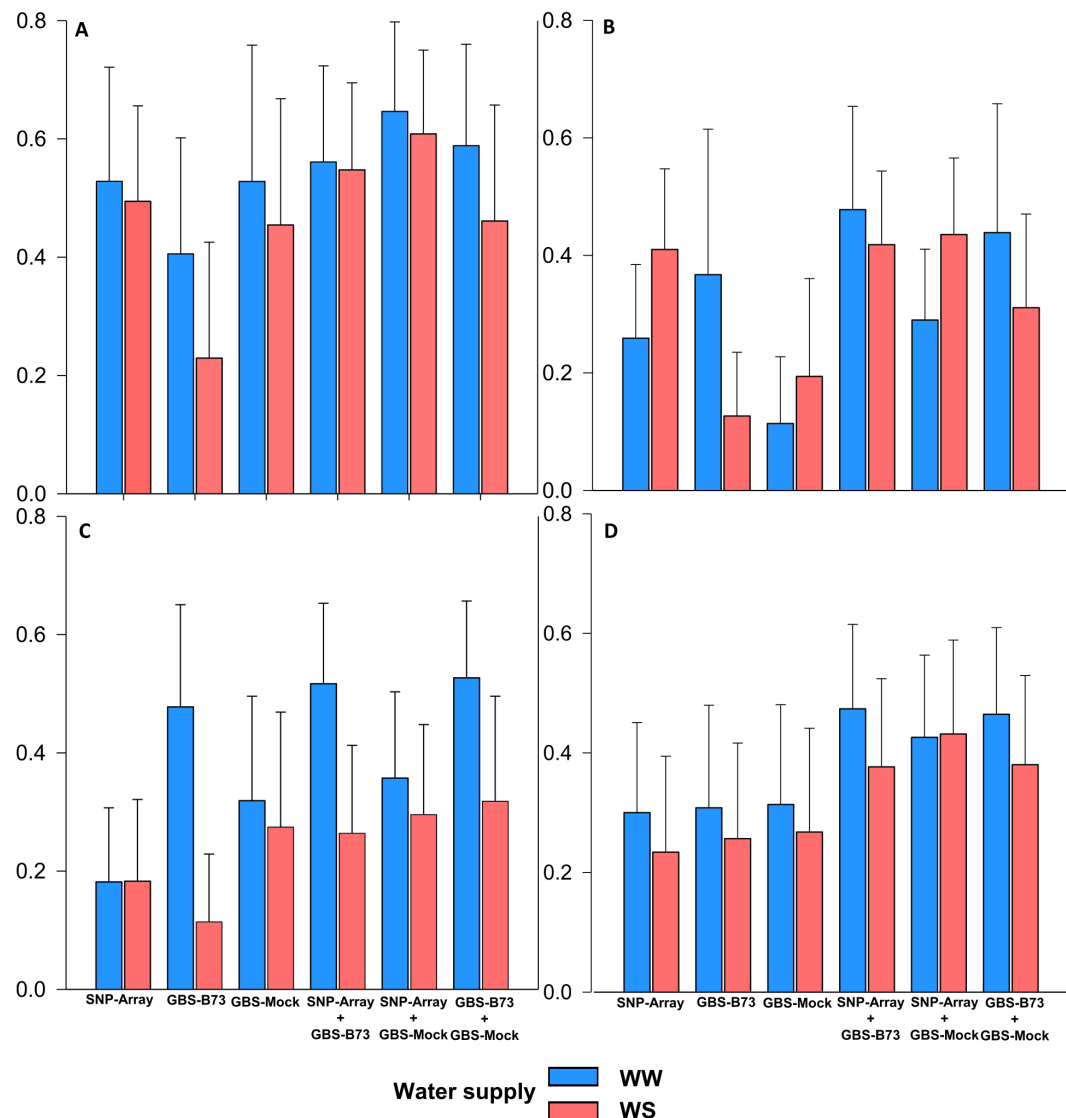


FIGURE 4

Proportion of phenotypic variance explained by the SNP (R^2_{OT}) per trait in WW (well-watered) and WS (water-stressed) conditions water supply and genotyping scenario. (A) SPAD, (B) PH (plant height), (C) SD (stalk diameter), and (D) SDM (shoot dry matter).

understanding genetic and phenotypic relationships, indicating that phenotypes are more interconnected than initially thought. This complicates the interpretation of study results, as it may need to be clarified which phenotype is directly influenced by the variant and to what extent. In genetic improvement studies, pleiotropic effects can affect the selection of desirable traits, as a single genetic variant can influence multiple agronomic or desirable traits.

The candidate gene *Zm00001d005090* is associated with SD under both water conditions and SDM under water deficit, possibly indicating a pleiotropic effect regulating the expression of these two traits. This gene is responsible for the clathrin heavy chain, one of the main subunits of clathrin, an essential protein in eukaryotic cells playing a crucial role in the endocytosis process. Hence, endocytosis occurs in many vital processes for plant development, such as abscisic acid (ABA) responses (Sutter et al., 2007). These authors state that in situations involving ABA, specific proteins in the

plasma membrane are negatively regulated through the induction of their endocytosis. It has been demonstrated that ABA and salicylic acid positively regulate a gene encoding a clathrin chain in maize (Zeng et al., 2013). ABA is produced in various parts of plants, including the stem, and it influences gene expression by activating stress-response protein-coding genes and repressing growth-related genes. There is also evidence that clathrin impacts Arabidopsis's stomatal function, gas exchange, and vegetative growth (Larson et al., 2017). Thus, this gene may have a pleiotropic effect, resulting in reduced height, stem diameter growth, and dry mass.

SNP were found to be associated with the trait simultaneously in both water availability levels, such as the gene *Zm00001d017978* identified in association with the PH trait in the SNP-Array scenario and the gene *Zm00001d001852* in association with the SD trait in the GBS-Mock scenario. *Zm00001d017978* has a putative function in the

endoglucanase enzyme, a subgroup of a larger enzyme family called cellulase. Cellulases are part of a superfamily of enzymes called hydrolases that use water to break down molecules. All cellulases are essential to degrading cellulose, a structural polysaccharide found in plant cell walls (Rahman et al., 2018). The cell wall plays a crucial role in plants' support and mechanical support, allowing them to grow by providing rigidity and resistance. Therefore, any alteration in cellulose degradation, caused by overexpression or underexpression of enzymes can affect structural integrity and, consequently, plant height. The applied water deficit may have negatively affected stem elongation, contributing to plant height, as at the V6 stage, the stem initiates the accelerated elongation phase. The gene *Zm00001d001852* has a putative function as Gibberellin-regulated protein 2 (GRP) with expression positively regulated by gibberellin. The plant hormone gibberellin regulates major aspects of plant growth and development (Yamaguchi, 2008), stimulating cell division and growth. The effect of gibberellin on stem diameter may be related to cell division and radial expansion of cells, increasing the number of cell layers. Additionally, there is evidence that biotic stresses impact gibberellin and GRP levels, as it has been reported that a slight increase in temperature can raise endogenous gibberellin concentration (Camut et al., 2019).

The genes *Zm00001d042735* and *Zm00001d031759* were also identified at both water supply levels and are associated with the SPAD trait. The first one was identified in the GBS-Mock scenario, while the other one was identified in both the SNP-Array and GBS-B73 scenarios, and both belong to the zinc finger family. Zinc finger proteins are named for their three-dimensional structure resembling a finger, binding to zinc ions through amino acids in the peptide sequence and widely distributed in eukaryotic organisms (Han et al., 2020). They bind to specific genetic sequences, interact with various proteins, participate in signal transduction, and regulate gene expression, playing an essential role in growth, development, and environmental adaptation. *Zm00001d042735* was described as a RING-type E3 ubiquitin transferase. Ubiquitin is a protein that acts as a molecular marker, signaling various cellular functions such as protein degradation, cell cycle regulation, cellular stress response, and intracellular signaling (Lee and Kim, 2011). E3 ubiquitin proteins respond to water stress by regulating ABA biosynthesis and signal transduction, modifying and degrading stress-related proteins (Han et al., 2022). An example is ZmAIRP4, which is involved in maize's ABA signaling, and this gene's overexpression increased water stress tolerance in *Arabidopsis* (Yang et al., 2018). Changes in water content induced by water stress can directly affect the SPAD index and chlorophyll content, as ABA concentration increases, causing stomatal closure to reduce water loss, which may affect the expression of genes related to stress response.

The gene *Zm00001d031759*, also belonging to the zinc finger protein family, has a putative function in the Protein shoot gravitropism 5 group, acting in the morphogenesis of aerial organs and responses to gravitropism. Some genes from the shoot gravitropism family have been identified and are involved in the perception and signal transduction for gravity associated with the branching angle (Yamauchi et al., 1997). It has also been found that loss of functionality of the shoot gravitropism five gene (SGR5) resulted in decreased starch accumulation in aerial tissues and consequently reduced gravity sensitivity (Tanimoto et al., 2008).

Gravity is an important regulator of plant architecture, allowing plants to optimize their position relative to the soil for nutrient absorption and to light for photosynthesis. Furthermore, some genes and regions manifest for the expression of the trait independently of the water supply level, probably unrelated to water stress.

4.2.1 Genes associated with phytohormone signaling pathway

Genes and regions shared among the genotyping scenarios were identified based on their function for the same trait (Supplementary Table S3). For example, genes *Zm00001d026477* in SNP-Array and *Zm00001d027695* in GBS-B73 are related to jasmonic acid (JA) response, associated with SPAD in WW traits. Jasmonate ZIM domain proteins, well-known as JAZ proteins, play a crucial role in pathogen responses (Ishiga et al., 2013) and are important signaling molecules in the JA pathway (Liu et al., 2017). Glutaredoxins are associated with water-induced stress response in maize, also participating in the abiotic stress response mediated by JA and ethylene through their interaction with transcription factors (Ding et al., 2019). As JA is involved in various signaling pathways regulating physiological and molecular processes in plants, in defense against biotic and abiotic stresses, such as drought (Rehman et al., 2023), signaling pathways induce stomatal closure, activating potassium efflux in guard cell protoplasts (Evans, 2003) enhancing the plants' ability to cope with environmental stresses.

Regarding ABA regulation, *Zm00001d016786* was associated with SDM in WS in SNP-Array and *Zm00001d005090* in GBS-B73. *Zm00001d021708* was found in GBS-B73, and *Zm00001d012719* in GBS-Mock for PH under WW conditions. *Zm00001d005090* and *Zm00001d053262* were also identified in GBS-B73 and GBS-Mock, respectively, for SD under WS conditions. Protein disulfide-isomerase (PDI) is a member of the thioredoxin superfamily of redox proteins with multiple physiological functions (Khan and Mutus, 2014), playing a crucial role in abiotic stress tolerance. Thioredoxin (TRXo1) is involved in ABA perception through redox regulation of specific receptors (De Brasi-Velasco et al., 2023). In maize, genes related to PDI were highly responsive to ABA and water stress (Liu et al., 2009). A PDI-like protein strongly associated with aboveground biomass and leaf size was also identified (Kang et al., 2015). According to Tanz et al. (2012), PDI is a family of proteins that affect chlorophyll biosynthesis in *Arabidopsis* seedlings.

The PPR (pentatricopeptide repeat) proteins are located in mitochondria or chloroplasts. In contrast, the BZIP (basic leucine zipper) proteins constitute a family of transcription factors (TFs) associated with plant growth, development, and stress responses. A typical PPR protein is targeted to mitochondria or chloroplasts, binds to one or several organellar transcripts, and influences their expression by altering RNA sequence, turnover, processing, or translation (Barkan and Small, 2014). It has been found that the PPR96 protein, located in mitochondria, altered the transcription levels of various stress-responsive genes under ABA treatments (Liu et al., 2016b). BZIP proteins are involved in multiple stress responses, primarily through the ABA signaling pathway (Uno et al., 2000). Changes in the transcription levels of maize BZIP TFs were observed in response to ABA treatments (Cao et al., 2019).

As mentioned earlier in SDM and SD, the Clathrin heavy chain indicates possible pleiotropy. Calcium-dependent lipid-binding proteins act in response to abiotic stress, such as drought. The expression of sANN3, a calcium-dependent lipid-protein, increased in response to water stress in rice, inducing various genes in the ABA signaling pathway and promoting root growth to enhance water absorption and stomatal closure to reduce water loss (Li et al., 2019). Therefore, these proteins and the biosynthesis pathways in ABA regulation may influence photosynthesis, plant development, and growth.

The genes *Zm00001d014899* in GBS-B73 and *Zm00001d001852* in GBS-Mock are associated with the trait SD under WW conditions, involved with the phytohormone gibberellin. The first encodes a protein from the tetratricopeptide repeat (TPR)-like superfamily. Proteins containing tetratricopeptide repeats play an important role in protein-protein interaction and regulating various cellular functions (Rosado et al., 2006). They serve different crucial roles in plants, including their involvement in phytohormone signaling, such as gibberellin (Jacobsen and Olszewski, 1993; Silverstone et al., 2007). Therefore, TPR-repeat-containing proteins are pivotal in signaling phytohormones and regulating various physiological processes, including growth, development, and environmental response. Gibberellin-regulated protein 2 (GRP) was mentioned earlier, occurring at both levels of water availability for SD.

Genes associated with SDM under WW conditions were found on the same chromosome, *Zm00001d031445* in the SNP-Array and *Zm00001d027626* in the GBS-B73, both involved in ethylene biosynthesis. The ethylene-insensitive3-like/ethylene-insensitive3 (EIL/EIN3) is one of the major regulatory families in ethylene signaling, also serving as a hub for ethylene connections with various plant responses to different environmental conditions (Wu and Yang, 2019). Ethylene is a crucial regulator in stress signaling, and its interaction with a receptor complex triggers the inactivation of kinase response, resulting in the initial dephosphorylation of EIN2, followed by the cleavage of the C-terminal of EIN2. Subsequently, EIN2 translocates to the nucleus, regulating the activation of EIN3/EIL1. These proteins, in turn, exert control over ethylene response factors (Yoshida et al., 2011).

S-adenosyl-L-methionine synthetase, known as SAM, is a donor of methyl groups in the biosynthesis of nucleic acids, proteins, lipids, polysaccharides, and secondary compounds (Heidari et al., 2020). SAM is involved in many important biological processes, such as ethylene biosynthesis. Yu et al. (2012) found that alterations in the expression level of SAM affected protein synthesis, phytohormones (JA and ethylene), and genes related to stress defense response. Ethylene is a volatile compound produced endogenously by plants for growth regulation - roots, stems, leaves, and flowers (Shilev, 2020). Plants increase the synthesis of this hormone when subjected to stressful situations, whether biotic or abiotic. Water deficit, in particular, is one of the main factors related to its increase (Apelbaum and Yang, 1981). Thus, the plant alters its growth rates, decreases biomass, and reduces development (Glick, 2014).

4.2.2 Genes associated with the circadian clock

Zm00001d044194 was identified in the SNP-Array, and *Zm00001d018127* in the GBS-B73 under WW condition associated with the SPAD trait acting in the circadian clock. The MYB proteins constitute one of the most extensive families of transcription factors found in plants, playing an important role in growth and development, with widespread expression in the development of corn and soybeans in stress responses, and are closely correlated with the circadian rhythm (Du et al., 2013). MYB-related genes can act as repressors and activators associated with the circadian clock (Kamioka et al., 2016; Hsu et al., 2013; Hu et al., 2024; Schaffer et al., 1998).

The SNW/Ski domain protein is involved in the post-transcriptional regulation of circadian clock genes. SkipP interacts with the serine/arginine-rich spliceosomal protein 45 (SR45) and controls the circadian cycle through alternative splicing of circadian clock genes under biotic stress conditions (Wang et al., 2012). The circadian clock in plants refers to an internal timing system on a cycle of approximately 24 hours that regulates plants' behavioral and physiological processes, including photosynthesis (Niwa et al., 2009). Likely, each guard cell maintains its circadian rhythm, and a clock controlling stomatal opening seems advantageous for the plant, helping prevent unnecessary water loss through transpiration (Dodd et al., 2005; Gorton et al., 1993). Thus, besides the environmental and internal factors that influence stomatal function, the circadian pattern in regulating stomatal movements is advantageous as it can enhance photosynthetic and water use efficiency.

4.2.3 Genes associated with ubiquitination regulation

The genes associated with the PH trait under WS conditions were *Zm00001d053809* in SNP-Array and *Zm00001d042481* in GBS-B73, which are related to the regulation of protein ubiquitination. Culins neddylation modulates the ubiquitin ligase activity of the complex, leading to increased ubiquitination and degradation of target proteins by the proteasome (Biswas et al., 2007; Mohanty et al., 2021; Pan et al., 2004). Neddylation is the post-translational protein modification most closely related to the regulation of protein ubiquitination (Rabut and Peter, 2008).

Ubiquitin thioesterases play a fundamental role in regulating the degradation of proteins marked with ubiquitin in plants. The ubiquitin system regulates virtually all aspects of cellular function (Ernst et al., 2013), which is important in controlling abiotic stress and processes that affect agronomic traits. For example, the ubiquitin-proteasome system is an essential pathway for protein degradation in plant growth and development (Linden and Callis, 2020). The ubiquitin-proteasome system is involved in regulating transcription responsive to ABA, allowing plants to respond to abiotic stresses such as drought (Dreher and Callis, 2007). Thus, ubiquitination affects gene expression or protein abundance to determine agronomic traits and stress control, enabling dynamic adjustments in physiological and biochemical responses contributing to plant survival and adaptation under adverse conditions.

4.2.4 Coincident genes among genotyping scenarios

Concerning the SNP-Array and GBS-B73 genotyping scenarios, these platforms are based on the same reference genome (B73) and are physically fixed, making it possible to determine the physical position of the marker in the genome. The coincidence between genes and regions on the same chromosome occurred only for *Zm00001d031445* in the SNP-Array and *Zm00001d027626* in the GBS-B73, both on chromosome 1. However, it was observed that, even though there was no coincidence regarding the physical position of the markers and chromosomes, there was still similarity regarding the gene functions.

Considering the three scenarios, when identifying the gene and region, it was observed that there was coincidence only for one marker in the SPAD trait under both irrigation conditions. However, possible coincidences were highlighted when deeper analyses were conducted regarding the gene function. Negro et al. (2019) concluded that GBS and SNP-Array were complementary for detecting QTLs in maize, marking different haplotypes. In a study performed in barley by Darrier et al. (2019), GBS and SNP-Array were shown to be efficient in accessing diversity. Still, they accessed different regions of the genome. These are methods that will capture different SNPs, there will be differences in position, density and distribution of the marks. However, even though they captured different regions, there was a positive correlation between the similarity matrices of both approaches. Thus, even when accessing different genome regions, these platforms demonstrate that they can be complementary. In the study, there was also a coincidence for the simulated genome, GBS-Mock, validating the complementarity for this scenario.

4.3 Association of markers in genotyping scenarios

The correlation between the markers in the SNP-Array and GBS-Mock scenarios provides information about the location of the markers on the chromosomes. Identifying a marker highly correlated with the GBS-Mock suggests that this marker is likely on a specific chromosome. The strength of the correlation between two markers is related to their physical proximity; the closer the markers are, the stronger the linkage disequilibrium (LD) (Myles et al., 2009). When markers are closer, they are more likely to be inherited together, leading to a stronger correlation between them. This is because when two markers are very close, they have fewer opportunities for recombination during meiosis, the process of gamete formation, which maintains stable combinations of adjacent alleles across generations. This information can be useful for guiding research and providing an initial direction for investigating the specific position of the marker in the genome.

However, according to the study results, the markers are located throughout the genome and not necessarily physically close. In other words, despite the relationship between the correlation's strength and the markers' physical proximity, the results showed that the markers are distributed across the entire genome. This suggests that other factors, besides physical proximity, may

influence the correlation between the markers, such as genetic inheritance patterns, recombination rate, and genomic structure, highlighting the importance of considering these aspects.

4.4 Combining genotyping scenarios

Combining genotyping scenarios can be a valid alternative for GWAS studies, providing higher resolution results than those obtained in isolated scenarios. In the approach involving Array and GBS, it was noticed that one tool complements the other, regardless of how GBS data are explored, whether with the referenced genome or *in-silico*, as there was little difference between SNP-Array + GBS-B73 and SNP-Array + GBS-Mock. Using multiple genotyping platforms, capturing a broader range of genetic markers in linkage disequilibrium with the loci of interest is possible. This can increase the ability to detect significant associations between genetic variants and phenotypes in GWAS studies.

Concerning the use of simulated genomes, MaChado et al. (2023) and Sabadin et al. (2022), assert that it is an excellent strategy for studies on diversity, population structure, heterotic group definition, tester selection, and genomic prediction for minor crops. Another caveat is that using temperate germplasm as a reference genome may introduce a significant bias when analyzing tropical germplasm (Xu et al., 2017). As a result, favorable alleles hidden in tropical maize, in specific tropical genomic regions, may be lost (Rasheed et al., 2017). With GBS, marker discovery and genotyping occur simultaneously, mitigating this bias and enabling the identification of markers in the analyzed diversity panel (Heslot et al., 2013). Furthermore, combining information obtained via conventional approaches with a reference genome obtained from the simulated genome should improve accuracy in association studies and impact the advancement of genetic research and the development of breeding strategies.

5 Final remarks

Negro et al. (2019) and Darrier et al. (2019) highlighted the complementarity between standard genotyping platforms for GWAS, demonstrating that both SNP-Array and GBS can identify markers strongly linked to genes influencing key phenotypic traits. However, adopting different genotyping platforms may incur substantial costs due to their distinct methodologies. Conversely, GBS genotyping offers the flexibility to utilize both the reference genome and *in-silico* genome, thereby avoiding additional expenses associated with combining these scenarios. In our study, combining GBS-B73 and GBS-Mock datasets resulted in a notable increase in accuracy for several traits compared to the highest accuracy achieved by GBS alone. Specifically, we observed accuracy gains of 0.06, 0.03, 0.05, and 0.15 for SPAD, PH, SD, and SDM, respectively. This integration of datasets allows for more comprehensive analyses, capturing a broader range of SNPs and providing enhanced resolution in explaining phenotypic variation. Ultimately, leveraging a single genotyping method enables more informative and efficient data exploration, facilitating a deeper understanding of the genetic basis of traits and informing crop improvement strategies.

Indeed, when a study aims to uncover greater genetic polymorphism within a species, and SNP-Array technology is unavailable, leveraging GBS approaches becomes a viable alternative. By conducting GWAS using GBS methods, researchers can effectively identify additional polymorphisms, thereby increasing the resolution and depth of the study. This strategy is particularly beneficial for minor or orphan crops with a genome reference but need access to SNP-Array technology. In such cases, GBS offers a cost-effective and accessible means to explore the genetic diversity present within these crops, facilitating a more comprehensive understanding of their genetic architecture and potential avenues for crop improvement. By harnessing the power of GBS-based GWAS, researchers can unlock valuable insights into the genetic factors underlying traits of interest, ultimately contributing to the development of improved varieties tailored to the specific needs of these crops.

Maize, with its high genetic diversity, can target pangenomes to improve the accuracy of genetic and phenotypic analyses (Lu et al., 2015). Pangenomes offer a more comprehensive representation of genomic variations within the species, allowing for the capture of rare or subpopulation-specific variations that may not be present in a single reference genome (Marschall et al., 2018). However, if the resources needed to generate or utilize pangenomes are not available, GBS remains an effective and accessible alternative. Compared to other approaches, GBS provides greater flexibility and sufficient resolution to identify significant polymorphisms, contributing to the exploration of genetic diversity and the advancement of breeding programs. By leveraging the power of GBS-based GWAS, either alongside pangenomes or as an alternative to them, researchers can unlock valuable insights into the genetic factors underlying traits of interest. This ultimately contributes to the development of improved varieties tailored to the specific needs of these crops.

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

FP: Data curation, Investigation, Methodology, Writing – original draft, Writing – review & editing. IM: Data curation, Writing – review & editing. MS: Data curation, Writing – review & editing. AL: Data curation, Writing – review & editing. FS: Data curation, Methodology, Validation, Writing – review & editing. RF-N:

Conceptualization, Formal Analysis, Investigation, Writing – review & editing, Data curation, Methodology. JD: Conceptualization, Data curation, Formal Analysis, Investigation, Writing – review & editing, Methodology.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was financially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)—Process: 409096/2021-0 CNPq/MCTI/FNDCT and by Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Funcap) - Process: 10504968/2023 - Edital 08/2023 – Funcap - Internacionalização.

Acknowledgments

We acknowledge the financial support provided by the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Finance Code 001, and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) - Process 2017/24327-0. We also extend our gratitude to Limagrain Field Seeds for their technical support. This manuscript was derived from the doctoral thesis of the first author.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could potentially create a 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.2024.1442008/full#supplementary-material>

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Apelbaum, A., and Yang, S. F. (1981). Biosynthesis of stress ethylene induced by water deficit. *Plant Physiol.* 68, 594–596. doi: 10.1104/pp.68.3.594
- Assareh, A., Volkert, L. G., and Li, J. (2012). Interaction Trees: Optimizing Ensembles of Decision Trees for Gene-Gene Interaction Detections. *2012 11th International Conference on Machine Learning and Applications*. Boca Raton, FL, USA. pp. 616–621. doi: 10.1109/ICMLA.2012.114
- Barkan, A., and Small, I. (2014). Pentatricopeptide repeat proteins in plants. *Annu. Rev. Plant Biol.* 65, 415–442. doi: 10.1146/annurev-arplant-050213-040159
- Biswas, K. K., Ooura, C., Higuchi, K., Miyazaki, Y., Van Nguyen, V., Rahman, A., et al. (2007). Genetic characterization of mutants resistant to the antiauxin p-chlorophenoxyisobutyric acid reveals that AAR3, a gene encoding a DCN1-like protein, regulates responses to the synthetic auxin 2,4-dichlorophenoxyacetic acid in arabidopsis roots. *Plant Physiol.* 145, 773–785. doi: 10.1104/pp.107.104844
- Bouchet, S., Bertin, P., Presterl, T., Jamin, P., Coubriche, D., Gouesnard, B., et al. (2017). Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. *Heredity (Edinb)*. 118, 249–259. doi: 10.1038/hdy.2016.88
- Browning, B. L., and Browning, S. R. (2008). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84, 210–223. doi: 10.1016/j.ajhg.2009.01.005
- Camut, L., Regnault, T., Sirlin-Josserand, M., Sakvelidze-Achard, L., Carrera, E., Zumsteg, J., et al. (2019). Root-derived GA12 contributes to temperature-induced shoot growth in Arabidopsis. *Nat. Plants* 5, 1216–1221. doi: 10.1038/s41477-019-0568-8
- Cao, L., Lu, X., Zhang, P., Wang, G., Wei, L., and Wang, T. (2019). Systematic analysis of differentially expressed maize ZmbZIP genes between drought and rewetting transcriptome reveals bZIP family members involved in abiotic stress responses. *Int. J. Mol. Sci.* 20. doi: 10.3390/ijms20174103
- Covarrubias-Pazarán, G. (2016). Genome-Assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11, 1–15. doi: 10.1371/journal.pone.0156744
- Da, Y., Wang, C., Wang, S., and Hu, G. (2014). Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One* 9. doi: 10.1371/journal.pone.0087666
- Darrier, B., Russell, J., Milner, S. G., Hedley, P. E., Shaw, P. D., Macaulay, M., et al. (2019). A comparison of mainstream genotyping platforms for the evaluation and use of barley genetic resources. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00544
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012
- De Brasi-Velasco, S., Sánchez-Guerrero, A., Castillo, M. C., Vertommen, D., León, J., Sevilla, F., et al. (2023). Thioredoxin TRXo1 is involved in ABA perception via PYR1 redox regulation. *Redox Biol.* 63. doi: 10.1016/j.redox.2023.102750
- De Souza Silveira, M. V., de Pontes, F. C. F., Machado, I. P., Lobo, A. L. A., Fritsche-Neto, R., and DoVale, J. C. (2024). Association mapping for image-based root traits in tropical maize under water stress in semi-arid regions. *Agron. J.* 116 (3), 1250–1264. doi: 10.1002/agg2.21528
- Ding, S., He, F., Tang, W., and Du, H. (2019). Identification of maize CC-type glutaredoxins that. *Genes (Basel)*. 10, 610. doi: 10.3390/g2Fgenes10080610
- Dodd, A. N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., et al. (2005). Cell biology: Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* 80(309), 630–633. doi: 10.1126/science.1115581
- Dreher, K., and Callis, J. (2007). Ubiquitin, hormones and biotic stress in plants. *Ann. Bot.* 99, 787–822. doi: 10.1093/aob/mcl255
- Du, H., Wang, Y., Xie, Y., Liang, Z., Jiang, S. J., Zhang, S. S., et al. (2013). Genome-wide identification and evolutionary and expression analyses of MYB-related genes in land plants. *DNA Res.* 20, 437–448. doi: 10.1093/dnares/dst021
- Elbasyoni, I. S., Lorenz, A. J., Guttieri, M., Frels, K., Baenziger, P. S., Poland, J., et al. (2018). A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. *Plant Sci.* 270, 123–130. doi: 10.1016/j.plantsci.2018.02.019
- Ernst, A., Avvakumov, G., Tong, J., Fan, Y., Zhao, Y., Alberts, P., et al. (2013). A strategy for modulation of enzymes in the ubiquitin system. *Sci. (80-)*. 339, 590–595. doi: 10.1126/science.1230161
- Evans, N. H. (2003). Modulation of guard cell plasma membrane potassium currents by methyl jasmonate. *Plant Physiol.* 131, 8–11. doi: 10.1104/pp.014266
- Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. 4.ed (London: Longman).
- Frensch, J., and Hsiao, T. C. (1994). Transient responses of cell turgor and growth of maize roots as affected by changes in water potential. *Plant Physiol.* 104, 247–254. doi: 10.1104/pp.104.1.247
- Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (zea mays L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6. doi: 10.1371/journal.pone.0028334
- Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., et al. (2014). TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. *PLoS One* 9. doi: 10.1371/journal.pone.0090346
- Glick, B. R. (2014). Bacteria with ACC deaminase can promote plant growth and help to feed the world. *Microbiol. Res.* 169, 30–39. doi: 10.1016/j.micres.2013.09.009
- Gorjanc, G., Bijma, P., and Hickey, J. M. (2015). Reliability of pedigree-based and genomic evaluations in selected populations. *Genet. Sel. Evol.* 47. doi: 10.1186/s12711-015-0145-1
- Gorton, H. L., Williams, W. E., and Assmann, S. M. (1993). Circadian rhythms in stomatal responsiveness to red and blue light. *Plant Physiol.* 103, 399–406. doi: 10.1104/pp.103.2.399
- Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., et al. (2020). C2H2 zinc finger proteins: master regulators of abiotic stress responses in plants. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00115
- Han, G., Qiao, Z., Li, Y., Yang, Z., Wang, C., Zhang, Y., et al. (2022). RING zinc finger proteins in plant abiotic stress tolerance. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.877011
- Heidari, P., Mazloomi, F., Nussbaumer, T., and Barcaccia, G. (2020). Insights into the SAM Synthetase Gene Family and Its Roles in Tomato Seedlings under Abiotic Stresses and Hormone Treatments. *Plants* 9, 586. doi: 10.3390/plants9050586
- Heslot, N., Rutkoski, J., Poland, J., Jannink, J. L., and Sorrells, M. E. (2013). Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. *PLoS One* 8. doi: 10.1371/journal.pone.0074612
- Hsiao, T. C. (1973). Plant responses to water stress. *Annu. Rev. Plant Physiol.* 24, 519–570. doi: 10.1146/annurev.pp.24.060173.002511
- Hsu, P. Y., Devisetty, U. K., and Harmer, S. L. (2013). Accurate timekeeping is controlled by a cycling activator in Arabidopsis. *Elife* 2013, 1–20. doi: 10.7554/eLife.00473
- Hu, Z., Zhang, N., Qin, Z., Li, J., Yang, N., Chen, Y., et al. (2024). Differential response of MYB transcription factor gene transcripts to circadian rhythm in tea plants (Camellia sinensis). *Int. J. Mol. Sci.* 25. doi: 10.3390/ijms25010657
- Ingvasson, P. K., and Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytol.* 189, 909–922. doi: 10.1111/j.1469-8137.2010.03593.x
- IPCC and Core Writing Team (2023). “Climate change 2023: synthesis report, summary for policymakers,” in *Contribution of working groups I, II and III to the sixth assessment report of the intergovernmental panel on climate change*. Eds. H. Lee and J. Romero (Geneva, Switzerland: IPC). doi: 10.59327/IPCC/AR6-9789291691647.001
- Ishiga, Y., Ishiga, T., Uppalapati, S. R., and Mysore, K. S. (2013). Jasmonate ZIM-domain (JAZ) protein regulates host and nonhost pathogen-induced cell death in tomato and nicotiana benthamiana. *PLoS One* 8. doi: 10.1371/journal.pone.0075728
- Jacobsen, S. E., and Olszewski, N. E. (1993). Mutations at the SPINDLY locus of Arabidopsis alter gibberellin signal transduction. *Plant Cell* 5, 887–896. doi: 10.1105/tpc.5.8.887
- Kamioka, M., Takao, S., Suzuki, T., Taki, K., Higashiyama, T., Kinoshita, T., et al. (2016). Direct repression of evening genes by CIRCADIAN CLOCK-ASSOCIATED1 in the Arabidopsis circadian clock. *Plant Cell* 28, 696–711. doi: 10.1105/tpc.15.00737
- Kang, Y., Sakiroglu, M., Krom, N., Stanton-Geddes, J., Wang, M., Lee, Y. C., et al. (2015). Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula. *Plant Cell Environ.* 38, 1997–2011. doi: 10.1111/pce.12520
- Khan, H. A., and Mutus, B. (2014). Protein disulfide isomerase a multifunctional protein with multiple physiological roles. *Front. Chem.* 2. doi: 10.3389/fchem.2014.00070
- Larcher, W. (2003). *Physiological plant ecology: ecophysiology and stress physiology of functional groups* (Berlin: Springer Nature).
- Larson, E. R., van Zelm, E., Roux, C., Marion-Poll, A., and Blatt, M. R. (2017). Clathrin heavy chain subunits coordinate endo- and exocytic traffic and affect stomatal movement. *Plant Physiol.* 175, 708–720. doi: 10.1104/pp.17.00970
- Lee, J. H., and Kim, W. T. (2011). Regulation of abiotic stress signal transduction by E3 ubiquitin ligases in Arabidopsis. *Mol. Cells* 31, 201–208. doi: 10.1007/s10059-011-0031-9
- 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, X., Zhang, Q., Yang, X., Han, J., and Zhu, Z. (2019). OsANN3, a calcium-dependent lipid binding annexin is a positive regulator of ABA-dependent stress tolerance in rice. *Plant Sci.* 284, 212–220. doi: 10.1016/j.plantsci.2019.04.019
- Linden, K. J., and Callis, J. (2020). The ubiquitin system affects agronomic plant traits. *J. Biol. Chem.* 295, 13940–13955. doi: 10.1074/jbc.REV120.011303
- Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016a). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* 12. doi: 10.1371/journal.pgen.1005767
- Liu, F., Sun, T., Wang, L., Su, W., Gao, S., Su, Y., et al. (2017). Plant jasmonate ZIM domain genes: Shedding light on structure and expression patterns of JAZ gene family in sugarcane. *BMC Genomics* 18, 1–17. doi: 10.1186/s12864-017-4142-3
- Liu, Y. H., Wang, X. T., Shi, Y. S., Huang, Y. Q., Song, Y. C., Wang, T. Y., et al. (2009). Expression and characterization of a protein disulfide isomerases in maize (Zea mays L.). *Chin. J. Biochem. Mol. Biol.* 25, 229–234.

- Liu, J. M., Zhao, J. Y., Lu, P. P., Chen, M., Guo, C. H., Xu, Z. S., et al. (2016b). The E-subgroup pentatricopeptide repeat protein family in *Arabidopsis thaliana* and confirmation of the responsiveness PPR96 to abiotic stresses. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01825
- Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., et al. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* 6. doi: 10.1038/ncomms7914
- MaChado, I. P., DoVale, J. C., Sabadin, F., and Fritsche-Neto, R. (2023). On the usefulness of mock genomes to define heterotic pools, testers, and hybrid predictions in orphan crops. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1164555
- Marschall, T., Marz, M., Abeel, T., Dijkstra, L., Dutilh, B. E., Ghaffari, A., et al. (2018). Computational pan-genomics: Status, promises and challenges. *Brief. Bioinform.* 19, 118–135. doi: 10.1093/bib/bbw089
- Melo, A. T. O., Bartaula, R., and Hale, I. (2016). GBS-SNP-CROP: A reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. *BMC Bioinf.* 17, 1–15. doi: 10.1186/s12859-016-0879-y
- Mohanty, P., Chatterjee, K. S., and Das, R. (2021). NEDD8 deamidation inhibits cullin RING ligase dynamics. *Front. Immunol.* 12. doi: 10.3389/fimmu.2021.695331
- Morosini, J. S., Mendonça, L., de, F., Lyra, D. H., Galli, G., Vidotti, M. S., et al. (2017). Association mapping for traits related to nitrogen use efficiency in tropical maize lines under field conditions. *Plant Soil* 421, 453–463. doi: 10.1007/s11104-017-3479-3
- Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., et al. (2009). Association mapping: Critical considerations shift from genotyping to experimental design. *Plant Cell* 21, 2194–2202. doi: 10.1105/tpc.109.068437
- Negro, S. S., Millet, E. J., Madur, D., Bauland, C., Combes, V., Welcker, C., et al. (2019). Genotyping-by-sequencing and SNP-arrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies. *BMC Plant Biol.* 19, 1–22. doi: 10.1186/s12870-019-1926-4
- Niwa, Y., Yamashino, T., and Mizuno, T. (2009). The Circadian Clock Regulates the Photoperiodic Response of Hypocotyl Elongation through a Coincidence Mechanism in *Arabidopsis thaliana*. *Plant Cell Physiol.* 50, 838–854. doi: 10.1093/pcp/pcp028
- Pan, Z.-Q., Kentsis, A., Dias, D. C., Yamoah, K., and Wu, K. (2004). Nedd8 on cullin: building an expressway to protein destruction. *Oncogene* 23, 1985–1997. doi: 10.1038/sj.onc.1207414
- Peterson, R. A. (2021). Estimating normalization transformations with bestNormalize. *The R Journal*. 13:1, 310–329. doi: 10.32614/RJ-2021-041
- Poland, J. A., Brown, P. J., Sorrells, M. E., and Jannink, J. L. (2012). Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7. doi: 10.1371/journal.pone.0032253
- Rabut, G., and Peter, M. (2008). Function and regulation of protein neddylation. “Protein modifications: beyond the usual suspects” review series. *EMBO Rep.* 9, 969–976. doi: 10.1038/embor.2008.183
- Rahman, M. S., Fernando, S., Ross, B., Wu, J., and Qin, W. (2018). Endoglucanase (EG) activity assays chapter 13. *Methods Mol. Biol.* 169–183. doi: 10.1007/978-1-4939-7877-9
- Rasheed, A., Hao, Y., Xia, X., Khan, A., Xu, Y., Varshney, R. K., et al. (2017). Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant* 10, 1047–1064. doi: 10.1016/j.molp.2017.06.008
- 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. doi: 10.3390/plants8020034
- Rehman, M., Saeed, M. S., Fan, X., Salam, A., Munir, R., Yasim, M. U., et al. (2023). The multifaceted role of jasmonic acid in plant stress mitigation: an overview. *Plants* 12. doi: 10.3390/plants12233982
- Rosado, A., Schapire, A. L., Bressan, R. A., Harfouche, A. L., Hasegawa, P. M., Valpuesta, V., et al. (2006). The *Arabidopsis* tetratricopeptide repeat-containing protein TTL1 is required for osmotic stress responses and abscisic acid sensitivity. *Plant Physiol.* 142, 1113–1126. doi: 10.1104/pp.106.085191
- Sabadin, F., Carvalho, H. F., Galli, G., and Fritsche-Neto, R. (2022). Population-tailored mock genome enables genomic studies in species without a reference genome. *Mol. Genet. Genomics* 297, 33–46. doi: 10.1007/s00438-021-01831-9
- Sabiel, S. A. I., Abdelmula, A. A., Bashir, E., Khan Baloch, S., I Sabiel, S. A., Abdelmula, A. A., et al. (2013). Genetic Variation of Plant Height and Stem Diameter Traits in Maize (*Zea mays* L.) under Drought Stress at Different Growth Stages. *J. Nat. Sci. Res.* 4, 116–122.
- Schaffer, R., Ramsay, N., Samach, A., Corden, S., Putterill, J., Carré, I. A., et al. (1998). The late elongated hypocotyl mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* 93, 1219–1229. doi: 10.1016/S0092-8674(00)81465-8
- 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. *Sci.* (80-). 326, 1112–1115. doi: 10.1126/science.1178534
- Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., et al. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat. Genet.* 44, 825–830. doi: 10.3390/app10207326
- Plant-growth-promoting bacteria mitigating soil salinity stress in plants. *Appl. Sci.* 10, 1–20.
- Silverstone, A. L., Tseng, T.-S., Swain, S. M., Dill, A., Jeong, S. Y., Olszewski, N. E., et al. (2007). Functional analysis of SPINDLY in gibberellin signaling in *Arabidopsis*. *Plant Physiol.* 143, 987–1000. doi: 10.1104/pp.106.091025
- Sutter, J. U., Sieben, C., Hartel, A., Eisenach, C., Thiel, G., and Blatt, M. R. (2007). Abscissic acid triggers the endocytosis of the *Arabidopsis* KAT1 K⁺ channel and its recycling to the plasma membrane. *Curr. Biol.* 17, 1396–1402. doi: 10.1016/j.cub.2007.07.020
- Taiz, L., Zeiger, E., Moller, I. M., and Murphy, A. (2015). *Plant physiology and development*. 6th Edn (Sunderland, Massachusetts, USA: Sinauer Associates Incorporated).
- Tanimoto, M., Tremblay, R., and Colasanti, J. (2008). Altered gravitropic response, amyloplast sedimentation and circumnutation in the *Arabidopsis* shoot gravitropism 5 mutant are associated with reduced starch levels. *Plant Mol. Biol.* 67, 57–69. doi: 10.1007/s11103-008-9301-0
- Tanz, S. K., Kilian, J., Johnsson, C., Apel, K., Small, I., Harter, K., et al. (2012). The SCO2 protein disulphide isomerase is required for thylakoid biogenesis and interacts with LCHB1 chlorophyll a/b binding proteins which affects chlorophyll biosynthesis in *Arabidopsis* seedlings. *Plant J.* 69, 743–754. doi: 10.1111/j.1365-3113X.2011.04833.x
- Thomson, M. J. (2014). High-throughput SNP genotyping to accelerate crop improvement. *Plant Breed. Biotechnol.* 2, 195–212. doi: 10.9787/pbb.2014.2.3.195
- Turner, N. C., and Jones, M. M. (1980). “Turgor maintenance by osmotic adjustment: A Review and evaluation,” in *Adaptation of plants to water and high temperatures stress*. Eds. N. C. Turner and P. J. Kramer (New York: Wiley & Sons), 87–103.
- Uno, Y., Furihata, T., Abe, H., Yoshida, R., Shinzaki, K., and Yamaguchi-Shinozaki, K. (2000). *Arabidopsis* basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11632–11637. doi: 10.1073/pnas.190309197
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Wang, X., Wu, F., Xie, Q., Wang, H., Wang, Y., Yue, Y., et al. (2012). SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in *Arabidopsis*. *Plant Cell* 24, 3278–3295. doi: 10.1105/tpc.112.100081
- Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., et al. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-73321-8
- Wickham, H. (2011). ggplot2. *WIREs Comput. Stat.* 3, 180–185. doi: 10.1002/wics.147
- Williams, E., Piepho, H. P., and Whitaker, D. (2011). Augmented p-rep designs. *Biometrical J.* 53, 19–27. doi: 10.1002/bimj.201000102
- Wu, Y. S., and Yang, C. Y. (2019). Ethylene-mediated signaling confers thermotolerance and regulates transcript levels of heat shock factors in rice seedlings under heat stress. *Bot. Stud.* 60. doi: 10.1186/s40529-019-0272-z
- Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., et al. (2017). Enhancing genetic gain in the era of molecular breeding. *J. Exp. Bot.* 68, 2641–2666. doi: 10.1093/jxb/erx135
- Yamaguchi, S. (2008). Gibberellin metabolism and its regulation. *Annu. Rev. Plant Biol.* 59, 225–251. doi: 10.1146/annurev.arplant.59.032607.092804
- Yamauchi, Y., Fukaki, H., Fujisawa, H., and Tasaka, M. (1997). Mutations in the SGR4, SGR5 and SGR6 loci of *Arabidopsis thaliana* alter the shoot gravitropism. *Plant Cell Physiol.* 38, 530–535. doi: 10.1093/oxfordjournals.pcp.a029201
- Yang, L., Wu, L., Chang, W., Li, Z., Miao, M., Li, Y., et al. (2018). Overexpression of the maize E3 ubiquitin ligase gene ZmAIRP4 enhances drought stress tolerance in *Arabidopsis*. *Plant Physiol. Biochem.* 123, 34–42. doi: 10.1016/j.plaphy.2017.11.017
- Yassue, R. M., Carvalho, H. F., Gevartosky, R., Sabadin, F., Souza, P. H., Bonatelli, M. L., et al. (2021). On the genetic architecture in a public tropical maize panel of the symbiosis between corn and plant growth-promoting bacteria aiming to improve plant resilience. *Mol. Breed.* 41. doi: 10.1007/s11032-021-01257-6
- Yassue, R. M., Galli, G., Chun-Peng, J. C., Fritsche-Neto, R., and Gota, M. (2023). Genome-wide association analysis of hyperspectral reflectance data to dissect the genetic architecture of growth-related traits in maize under plant growth-promoting bacteria inoculation. *Plant Direct* 7. doi: 10.1002/pld3.492
- Yin, L. (2020). CMplot: circle manhattan plot. *C. Circ. Manhattan Plot*.
- Yoshida, T., Ohama, N., Nakajima, J., Kidokoro, S., Mizoi, J., Nakashima, K., et al. (2011). *Arabidopsis* HsfA1 transcription factors function as the main positive regulators in heat shock-responsive gene expression. *Mol. Genet. Genomics* 286, 321–332. doi: 10.1007/s00438-011-0647-7
- Yu, J. G., Lee, G. H., and Park, Y. D. (2012). Physiological role of endogenous S-adenosyl-L-methionine synthetase in Chinese cabbage. *Hortic. Environ. Biotechnol.* 53, 247–255. doi: 10.1007/s13580-012-0021-7
- Zeng, M. H., Liu, S. H., Yang, M. X., Zhang, Y. J., Liang, J. Y., Wan, X. R., et al. (2013). Characterization of a gene encoding clathrin heavy chain in maize up-regulated by salicylic acid, abscisic acid and high boron supply. *Int. J. Mol. Sci.* 14, 15179–15198. doi: 10.3390/ijms140715179
- Zhang, X., Ren, Z., Luo, B., Zhong, H., Ma, P., Zhang, H., et al. (2022). Genetic architecture of maize yield traits dissected by QTL mapping and GWAS in maize. *Crop J.* 10, 436–446. doi: 10.1016/j.cj.2021.07.008
- Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., and Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28, 3326–3328. doi: 10.1093/bioinformatics/bts606



OPEN ACCESS

EDITED BY

Robert Henry,
The University of Queensland, Australia

REVIEWED BY

Tian Tang,
Sun Yat-sen University, China
Zhaokui Du,
Taizhou University, China

*CORRESPONDENCE

Yan Zhu

✉ zhu_yan@fudan.edu.cn

Peng Nan

✉ nannpeng@fudan.edu.cn

RECEIVED 17 October 2024

ACCEPTED 27 February 2025

PUBLISHED 26 March 2025

CITATION

Zhang J, Ouyang S, Cai X, Yang S, Chen Q,
Yang J, Song Z, Zhang W, Wang Y, Zhu Y and
Nan P (2025) Phenotypic adaptation and
genomic variation of *Kandelia obovata*
associated with its northern introduction
along southeastern coast of China.
Front. Plant Sci. 16:1512620.
doi: 10.3389/fpls.2025.1512620

COPYRIGHT

© 2025 Zhang, Ouyang, Cai, Yang, Chen, Yang,
Song, Zhang, Wang, Zhu and Nan. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Phenotypic adaptation and genomic variation of *Kandelia obovata* associated with its northern introduction along southeastern coast of China

Jiaqi Zhang^{1,2}, Sheng Ouyang^{1,2}, Xingxing Cai^{1,2}, Sheng Yang³,
Qiuxia Chen³, Ji Yang^{1,2}, Zhiping Song^{1,2}, Wenju Zhang^{1,2},
Yuguo Wang^{1,2}, Yan Zhu^{1*} and Peng Nan^{1,2*}

¹School of Life Sciences, Fudan University, Shanghai, China, ²State Key Laboratory of Wetland Conservation and Restoration, National Observations and Research Station for Wetland Ecosystems of the Yangtze Estuary, Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, and Institute of Eco-Chongming, Shanghai, China, ³Wenzhou Key Laboratory of Resource Plant Innovation and Utilization, Zhejiang Institute of Subtropical Crops, Zhejiang Academy of Agricultural Sciences, Wenzhou, Zhejiang, China

Introduction: Mangroves play a crucial role within coastal wetland ecosystems, with *Kandelia obovata* frequently utilized for introduction studies and cultivation research. Investigating the rapid adaptability of *K. obovata* across diverse environmental conditions offers valuable insights into how mangroves can effectively acclimate to global climate fluctuations.

Methods: In this study, following a common garden experiment, we investigated variations in morphological traits between two distinct populations of *K. obovata*, Quanzhou (QZ) and Wenzhou (WZ), originating from the same introduction site Zhangzhou (ZZ). Then we performed the whole-genome resequencing on multiple populations along the southern coast of China to assess genetic divergence and diversity patterns in response to environmental factors.

Results: Our findings have uncovered divergent growth-defense trade-off mechanisms employed by these two populations when exposed to varying minimal temperatures in the coldest month within their respective habitats. Moreover, our observations have revealed discernible genetic divergence during the process of environmental acclimatization. Subsequent whole-genome re-sequencing have unveiled a significant decrease in genetic diversity within the northernmost population, suggesting that temperature plays a primary role in shaping genetic variability within the *K. obovata* species.

Discussion: These findings present new evidence for the rapid adaptation of *K. obovata* and contributes to our understanding of environmental adaptation characteristics during its introduction to northern regions, which holds significant implications for the conservation and sustainable development of mangroves.

KEYWORDS

Kandelia obovata, phenotypic adaption, genomic variation, northern introduction, common garden experiment

1 Introduction

Given the escalating severity of global climate change, understanding the mechanisms by which organisms adapt to environmental fluctuations has emerged as a pivotal area of scientific investigation in evolutionary biology (Bjorkman et al., 2017). The adaptability of plants stands as a critical determinant influencing their survival capacity within an ever-changing climatic milieu (Foyer and Kranner, 2023). Exploring the adaptive responses exhibited by plants towards diverse environmental alterations can provide valuable insights into their potential coping strategies amidst global climate change. Mangroves represent arboreal ecosystems thriving in tropical and subtropical intertidal zones, characterized by remarkable biodiversity and recognized as one of the foremost carbon-sequestering habitats worldwide. They play indispensable roles in storm resistance, coastal protection, and preservation of coastal ecological equilibrium (Li et al., 2024). Temperature serves as the primary environmental factor delimiting the geographical distribution of mangroves, making them among the most vulnerable groups affected by global climate change (Chen et al., 2017). Consequently, mangrove plants offer ideal experimental subjects for investigating rapid plant responses to global climate change. However, due to intensified human activities and global climate change, mangrove forests are experiencing considerable decline worldwide. Consequently, artificial plantation of mangroves has emerged as an essential strategy for ecosystem restoration (Ellison et al., 2020). There is abundant evidence suggesting that introduced mangrove forests in new environments often undergo adaptive evolution to cope with selection pressures specific to these novel habitats (Miriyeganeh, 2022).

Plants adopt phenotypic plasticity and genetic differentiation as two distinct strategies to adapt to their environment (Boquete et al., 2021; Bastias et al., 2024). During the course of evolution, plants must balance these two approaches. Initially, a plant species may utilize plasticity to mitigate natural selection pressure when colonizing a new environment. However, with time and expansion of its geographic range, plant populations may undergo genetic differentiation. In this process, plants can develop various ecotypic traits to adapt to diverse climatic and geographical conditions (Baythavong and Stanton, 2010). Studies have demonstrated that successful invasion by *Brachypodium sylvaticum* is driven by genetic variation rather than phenotypic plasticity induced by the environment (Marchini et al., 2019). A common garden experiment (CGE) can eliminate the influence of environmental factors from provenances on results and assess whether plant populations from different locations have genetic differentiation (Zhao et al., 2021).

Research has revealed that *K. obovata*, a member of the Rhizophoraceae family, is a widely distributed mangrove species in China renowned for its ability to thrive in high latitudes and withstand cold temperatures (Su et al., 2019). Previous study has indicated the presence of significant genetic differentiation and variation within *K. obovata* populations along the southeastern coast of China (Zhao et al., 2024). These populations exhibit substantial genetic diversity and may serve as key centers of

diversity for Asian mangroves. The genetic diversity of mangroves has been extensively investigated using molecular marker techniques such as Sequence-Related Amplified Polymorphism (SRAP) and Simple Sequence Repeats (SSR) in previous studies (Lu et al., 2021). However, with the rapid advancement of sequencing technology, a chromosome-level reference genome for *K. obovata* was published in 2020 (Hu et al., 2020). By employing resequencing and bioinformatics methods, studying the genomic level differences and adaptive evolution between populations enables us to comprehend the genetic mechanisms underlying organisms' adaptation to selection and identify crucial candidate genes. This provides a theoretical foundation for the rational utilization of germplasm resources (Lu et al., 2023).

As part of the exploration work on mangrove protection and introduction in China, *K. obovata* was introduced from the Jiulong River Estuary Mangrove Provincial Nature Reserve in Zhangzhou city (ZZ), Fujian Province (117°92'E, 24°46'N) to the Luoyang River Mangrove Nature Reserve in Quanzhou Bay, Quanzhou city (QZ), Fujian Province (118°59'E, 24°59'N) in 2003. Subsequently, it was further introduced to Longgang Aojiang Estuary Mangrove, Wenzhou city (WZ), Zhejiang Province (120°96'E, 28°12'N) in 2005. After approximately two decades since its original habitat to the introduction sites, significant phenotypic variations have been observed in these two populations of *K. obovata*. This study aims to investigate whether these differences are associated with genetic background through CGE. Additionally, using whole-genome re-sequencing (WGRS) technology, this study analyzes and compares the levels of genetic diversity between populations of *K. obovata* from its original habitat and introduction sites to unravel their phylogenetic relationships and genetic structure. These findings will provide genetic data support for identifying and utilizing germplasm resources of *K. obovata*. Furthermore, this study explores candidate genes under strong selective pressure across different geographical environments and predicts cis-regulatory elements within their promoter regions to elucidate the impact of environmental adaptation on population differentiation of *K. obovata*. This provides a theoretical basis for understanding rapid ecological adaptation mechanisms of *K. obovata* towards environmental changes and strengthens ecological research on global climate response and adaptability of mangroves. It is also significant for mangrove conservation, new variety breeding, and serves as a reference for studying mechanisms by which plants rapidly adapt to environmental changes.

2 Materials and methods

2.1 Common garden experiment

In May 2019, we acquired approximately 20,000 hypocotyls of *K. obovata* from two distinct locations: the Luoyang River Mangrove Nature Reserve in Quanzhou Bay, Quanzhou City (QZ), Fujian Province, and the Longgang Aojiang Estuary Mangrove in Wenzhou City (WZ), Zhejiang Province. These individuals were then transplanted into plastic pots at the Mangrove Base of Nanhui

Dongtan, Shanghai (SH) (121°97'E, 30°90'N) for cultivation under CGE conditions. The objective was to assess the overwintering survival rate of *K. obovata* seedlings under natural conditions (Supplementary Figure 1). In May 2021, we additionally acquired mature hypocotyls of *K. obovata* from both QZ and WZ. We conducted a random sampling of 30 healthy hypocotyls from each population and recorded their hypocotyl length (HL), hypocotyl weight (HW), and hypocotyl diameter (HD). Subsequently, we carefully selected 1,000 individuals with comparable weights, lengths, and sizes from each population. These individuals were also transplanted into plastic pots in SH for CGE cultivation, with the purpose of subsequent phenotypic observations. The Nanhui Dongtan, located in the Yangtze estuary, is the largest and most extensive coastal wetland in the region. It experiences a subtropical monsoon climate characterized by moderate temperatures and high humidity levels. This area exhibits features of both monsoon and maritime climates. With an average annual temperature ranging from 15 to 16 °C, it has recorded its highest temperature at 37.3 °C and lowest temperature at -7.9 °C.

2.2 Determination of morphological traits

After 18 months of growth in CGE, a total of 30 robust and healthy plants were randomly selected from both the QZ and WZ populations for comprehensive measurement of plant morphological traits. The assessment covered the following parameters: plant height (PH), basal diameter (BD), crown width (W), leaf number (LN), and branch number (BN). Thirty intact leaves were carefully collected from the upper branches of both QZ and WZ populations, thoroughly rinsed with tap water, and then precisely weighed to determine their fresh weight (LFW). After being soaked in distilled water for 12 hours followed by thorough drying to eliminate surface moisture, the leaves were reweighed to obtain their saturated fresh weight (LSFW). Subsequently, scanned images of the collected leaves were analyzed using ImageJ software (1.53t) (Schneider et al., 2012) to calculate various parameters such as leaf length (LL), leaf width (LW), leaf perimeter (LP), leaf area (LA), leaf shape index (LSI = LL/LW). The collected samples were placed in trays and subjected to constant drying at 60°C for 48 hours until reaching a stable weight. The dry weight (DW) was measured to compute relative water content (RWC = (LFW - DW)/(LSFW - DW)) and leaf dry matter content (LDMC = DW/LSFW). Specific leaf area (SLA) is calculated as LA/DW.

2.3 Leaf anatomical structure

We employed the Safranin-O/Fast green staining method to prepare *K. obovata* samples. Paraffin sections were made with the second leaf from the top down, and ImageJ software (1.53t) was utilized for quantifying the thickness of various leaf components: leaf thickness (Lt), upper cuticle layer thickness (UCu), upper epidermis thickness (UEp), upper hypodermis thickness (UHy), upper palisade tissue thickness (UPt), spongy tissue thickness (St), lower cuticle layer thickness (LCu), lower epidermis thickness (LEp), lower hypodermis thickness (LHy) as well as lower

palisade tissue thickness (LPt). The ratio of palisade to spongy tissues (P/S = Pt/St) was calculated along with cell tenses ratio (CTR = Pt/LT) and spongy ratio (SR = St/LT).

2.4 Assessment of stomatal characteristics

30 leaf samples were selected from both the QZ and WZ populations in CGE. Temporary slides of the lower epidermis stomata of *K. obovata* leaves were prepared using the nail polish imprinting method (Pathoumthong et al., 2023), observed, and photographed under a 10x optical microscope. ImageJ software (1.53t) was utilized to quantify the stomata number (SN), stomatal area (SA), and stomatal density (SD) per view field. 10-15 view fields were examined in each sample, and the results were averaged.

2.5 Determination of cold resistance

We measured the daily minimum temperature changes at Nanhui Dongtan Mangrove Base in Shanghai during the coldest months of 2020 and 2021 as shown in Supplementary Figure 2. Despite a warm winter in Shanghai in 2020, an extended period of extreme low temperatures (-7°C) occurred from December 29, 2020 to January 2, 2021. The overwinter survival rates of mangrove plants in their natural environment at the base were separately recorded on January 21, 2020 (without extreme low temperatures) and January 5, 2021 (during the cold wave with extreme low temperatures). The data was analyzed using Microsoft Excel (2021) software. Non-paired Student's *t*-test was employed to compare significant differences among different populations in various indicator values. Additionally, a Chi-square test with Yates' continuity correction was used to assess the differences in overwinter survival rates between the WZ and QZ populations under different temperature conditions. A significance level of $p > 0.05$ indicated non-significance; a range of $0.01 < p < 0.05$ denoted significance; while $p < 0.01$ represented high significance levels. R (4.2.2), Rstudio (2023.03.0 + 386) ggplot2 package and GraphPad Prism (8) software were utilized for graphical representation purposes.

2.6 Whole-genome re-sequencing

In mid to late July 2023, mature *K. obovata* leaves samples were collected from four locations: Dongzhaigang Mangrove Nature Reserve (DZG) in Haikou City, Hainan Province (110°58'E, 19° 95'N); Jiulong River Estuary Mangrove Provincial Nature Reserve in Zhangzhou City (ZZ), Fujian Province (117°92'E, 24°46'N); Luoyang River Mangrove Nature Reserve in Quanzhou Bay, Quanzhou City (QZ), Fujian Province (118°59'E, 24°91'N); and Longgang Aojiang Estuary Mangrove in Wenzhou City (WZ), Zhejiang Province (120°96'E, 28°12'N) (Supplementary Figure 1). A total of 40 samples were collected, comprising 10 mature *K. obovata* leaves from each location, with a minimum distance of over 50 meters maintained between the collection points.

The CTAB method was employed to extract DNA from the leaf samples. Only high-quality DNA samples ($OD_{260/280} = 1.8\sim 2.0$, $OD_{260/230} = 2.0$) were utilized for constructing the sequencing library. A total of 0.5 μ g of DNA per sample served as input material for the preparation of the DNA library. The Truseq Nano DNA HT Sample Prep Kit (Illumina USA) was used to generate the sequencing library in accordance with the manufacturer's recommendations, and index codes were assigned to each sample. In brief, genomic DNA samples were sonicated to obtain fragments with a size of 350 bp, followed by end-polishing, A-tailing, and ligation with full-length adapters suitable for Illumina sequencing technology; this was succeeded by additional PCR amplification steps. After purification of PCR products using the AMPure XP system, libraries underwent size distribution analysis via Agilent 2100 Bioanalyzer and quantification through real-time PCR (3nM). Finally, paired-end DNA-seq sequencing libraries were sequenced on an Illumina NovaSeq system at Shanghai Majorbio Bio-pharm Technology Co., Ltd.

2.7 Variant discovery

The raw reads of low quality (mean phred score < 20), which included reads containing adapter contamination and unrecognizable nucleotides (N base > 10), were trimmed or discarded using Fastp software (0.23.2) (Chen et al., 2018). After trimming, the reads were mapped to the reference genome (<https://bigd.big.ac.cn/gwh/Assembly/990/show>) using BWAMEME software (1.0.5) (Jung and Han, 2022) with default mapping parameters. The alignment bam files were sorted by SAMtools (1.15.1) (Li et al., 2009) and PCR duplicates were marked using MarkDuplicated as part of the modified GATK Best Practice pipeline (4.3.0.0) (McKenna et al., 2010). Base quality recalibration was performed, followed by germline variant calling for Single Nucleotide Polymorphisms (SNPs) across all samples using Haplotyper and Gvcftyper programs in Sentieon genomics tools (202112.07) (Freed et al., 2017). Variants were filtered according to standard hard filtering parameters based on GATK Best Practices pipeline (4.3.0.0), and annotated using SnpEff (5.1d) (Cingolani et al., 2012). Subsequently, several filtering steps were applied to reduce false positives for SNPs and genotype calling using VCFtools software (0.1.16) (Danecek et al., 2011): (i) SNPs with more than two alleles were removed, (ii) SNPs with mean depth values less than 4 across all samples were removed, (iii) SNPs with minor allele frequency < 0.05 were removed, (iv) Only SNPs that could be genotyped in at least 70% of the samples were retained, (v) For population structure analysis, SNPs showing linkage disequilibrium patterns were pruned using Plink software (1.90b6.20) (Purcell et al., 2007).

2.8 Genetic diversity analysis

Based on filtered vs. files, we calculated observed heterozygosity (Ho), expected heterozygosity (He) and nucleotide diversity (π)

using the populations module in the software Stacks (2.64) (Catchen et al., 2013). Additionally, GenAlEx (6.5) (Peakall and Smouse, 2012) was employed to determine genetic diversity parameters such as Polymorphic Information Content (PIC) and Shannon's Information Index. These parameters were utilized to evaluate the level of genetic diversity in four populations of *K. obovata* studied here. A higher value for these genetic diversity parameters indicates a greater level of genetic diversity within the population.

The Genetic Differentiation Index (F_{ST}) between populations was calculated using the populations module in the Stacks software (2.64) (Catchen et al., 2013). F_{ST} values were obtained through pairwise comparisons among all populations to evaluate the extent of genetic differentiation between two populations. The level of genetic differentiation can be determined based on a range of F_{ST} values: 0~0.05 indicates an extremely low degree of genetic differentiation between populations; 0.05~0.15 suggests a moderate level of genetic differentiation; 0.15~0.25 signifies a considerable degree of genetic differentiation between populations; and when $F_{ST} > 0.25$, it indicates a high level of genetic differentiation among populations (Wright, 1978).

2.9 Phylogenetic analyses

The Maximum Likelihood phylogenetic tree was constructed using IQ-TREE2 software (2.1.2) (Minh et al., 2020). The ML analyses were performed on the pruned SNP sites employing IQ-TREE2 with GTR+I+G4 model and 1000 bootstraps.

The unsupervised maximum-likelihood clustering algorithm implemented in ADMIXTURE (1.3.0) (Alexander et al., 2009) was used to cluster each genome in the investigated populations of *K. obovata*. Initial clustering was performed for $K = 1$ to $K = 20$ ancestral clusters using default settings. The optimal value of K was determined based on the cross-validation error (CV error), and the genetic structure corresponding to this optimal value of K was outputted as the final result. To enhance the accuracy of initial clustering, pruned SNPs were utilized in the structural analysis.

To visualize the genetic relationships among samples, we conducted principal component analysis (PCA) based on pruned SNPs using Plink (1.90b6.20) (Purcell et al., 2007).

2.10 Screening for selective sweep and identification of positively selected genes

Following the sliding window strategy, we utilized PIXY software (1.2.7.beta1) (Korunes and Samuk, 2021) to perform segmentation and calculation of F_{ST} as well as π between QZ and WZ populations. The window size was set at 10 Mb with a step size of 10 kb. To strengthen the analysis of positive selection, we employed a permutation test to account for random differentiation that may occur between populations in the absence of selection pressure. Specifically, we performed 1,000 permutation tests by randomly shuffling the population labels and recalculating

the F_{ST} distribution for each 10 kb window. This approach allowed us to simulate different population structures under the null hypothesis of no selection. The 95th percentile of the permutation distribution was used as a baseline to determine whether the observed F_{ST} values were significantly higher than expected by chance. Finally, we employed a screening criterion where regions with F_{ST} values in the top 5% were identified as potential candidate regions distinguishing QZ from WZ populations. Furthermore, our selective sweep analysis combined both F_{ST} and π values. Specifically, regions of the genome exceeding the top 5% threshold, along with regions exhibiting extremely high π ratios (π_{QZ}/π_{WZ}), were considered as potential areas displaying strong signals of selection scanning for WZ population. Sequence Toolkit in TBtools (2.112) (Chen et al., 2020) was used to associate the selected region with candidate genes to obtain the positively selected genes of WZ population.

The TBtools software (2.112) was also employed to conduct Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis on the identified regions from the QZ and WZ populations. Significance of enrichment was determined when p -value < 0.05 for both GO and KEGG pathways. Visualization was performed using the qqman and ggplot2 packages in R (4.2.2) and Rstudio (2023.03.0 + 386) software.

2.11 Analysis of cis-element in promoters

After extracting the 2,000 bp upstream sequence of the positive selection gene in the WZ population using TBtools software (2.112), we conducted promoter cis-element analysis through the PlantCare database (Lescot et al., 2002).

2.12 Integrated analysis of transcriptome and genome data

We retrieved previously generated transcriptome data (Zhang et al., 2024), and intersected the differentially expressed genes (DEGs) under cold stress from both populations with 40 candidate positively selected genes identified from the genome based on F_{ST} and π values. These genes were further analyzed for transcriptome expression profiles and promoter elements. Heatmap visualization was performed using ggplot2 and pheatmap packages in R (4.2.2) and RStudio (2023.03.0 + 386). Transcriptional levels and cis-elements were visualized using TBtools Simple BioSequence Viewer (2.112) and GraphPad Prism (8).

3 Results

3.1 Morphological differences between two populations of *K. obovata*

In comparison to the QZ population, the WZ population's hypocotyls exhibit relatively lower values for physiological

parameters (Supplementary Table 1). There are statistically significant differences ($p < 0.05$, non-paired Student's t -test) in HL, HW, HD, PH, BD, LN, and BN, as well as W between these two populations. Moreover, significant differences between the two populations ($p < 0.05$, non-paired Student's t -test) are also observed in SLA, and DW (Figure 1). However, no notable differences are found in LW, LSI, LA, RWC or LDMC.

3.2 Disparities in leaf anatomical structure and stomatal characteristics between two *K. obovata* Populations

The cross-sectional images of *K. obovata* reveal a leaf structure comprising upper and lower cuticles, epidermis, hypodermis, palisade tissue, and spongy tissue. Both the upper and lower epidermis consist of a single layer of cells, while the hypodermis consists of two cell layers—the first layer being smaller without tannins and the second layer larger containing tannins. The densely arranged palisade tissue has a thicker upper layer than the lower layer. The spongy tissue, which accommodates vascular bundles, displays a looser arrangement (Figures 1A, B). Furthermore, the Lt, St, and SR are significantly greater in the QZ population compared to those in the WZ population ($p < 0.05$, non-paired Student's t -test). Conversely, the WZ population exhibits significantly higher thickness in UHy, UPt, P/S, and CTR compared to the QZ population ($p < 0.05$, non-paired Student's t -test). Stomata of *K. obovata* are predominantly distributed in the lower epidermis (Figure 1B). The SA is significantly smaller in QZ population than that in WZ population ($p < 0.05$, non-paired Student's t -test), while SD is significantly larger than that in WZ population ($p < 0.05$, non-paired Student's t -test) (Figure 1A).

3.3 Variation in Overwintering Survival Rate between Two *K. obovata* Populations under Natural Conditions

One of the primary constraints on the northward expansion of *K. obovata* is its susceptibility to cold temperatures. We investigated the overwinter survival rates of both populations under natural conditions following the CGE (Table 1). In January 2020 with moderate winter temperatures, the survival rate of *K. obovata* was notably high. Specifically, the natural overwinter survival rate of WZ population was 84.5%, surpassing that of the QZ population (67.3%). Both populations exhibited reduced survival rates under extreme low temperatures in 2021; however, even under such conditions, the WZ population demonstrated a higher winter survival rate (18.6%) compared to that of the QZ population (6.8%). Our findings indicate a significant difference in overwinter survival rates between the two populations both under extreme low temperatures and moderate winter conditions ($p < 2.2e-16$, Chi-square test). The WZ population exhibits superior cold tolerance and better adaptation to the northern CGE environment.

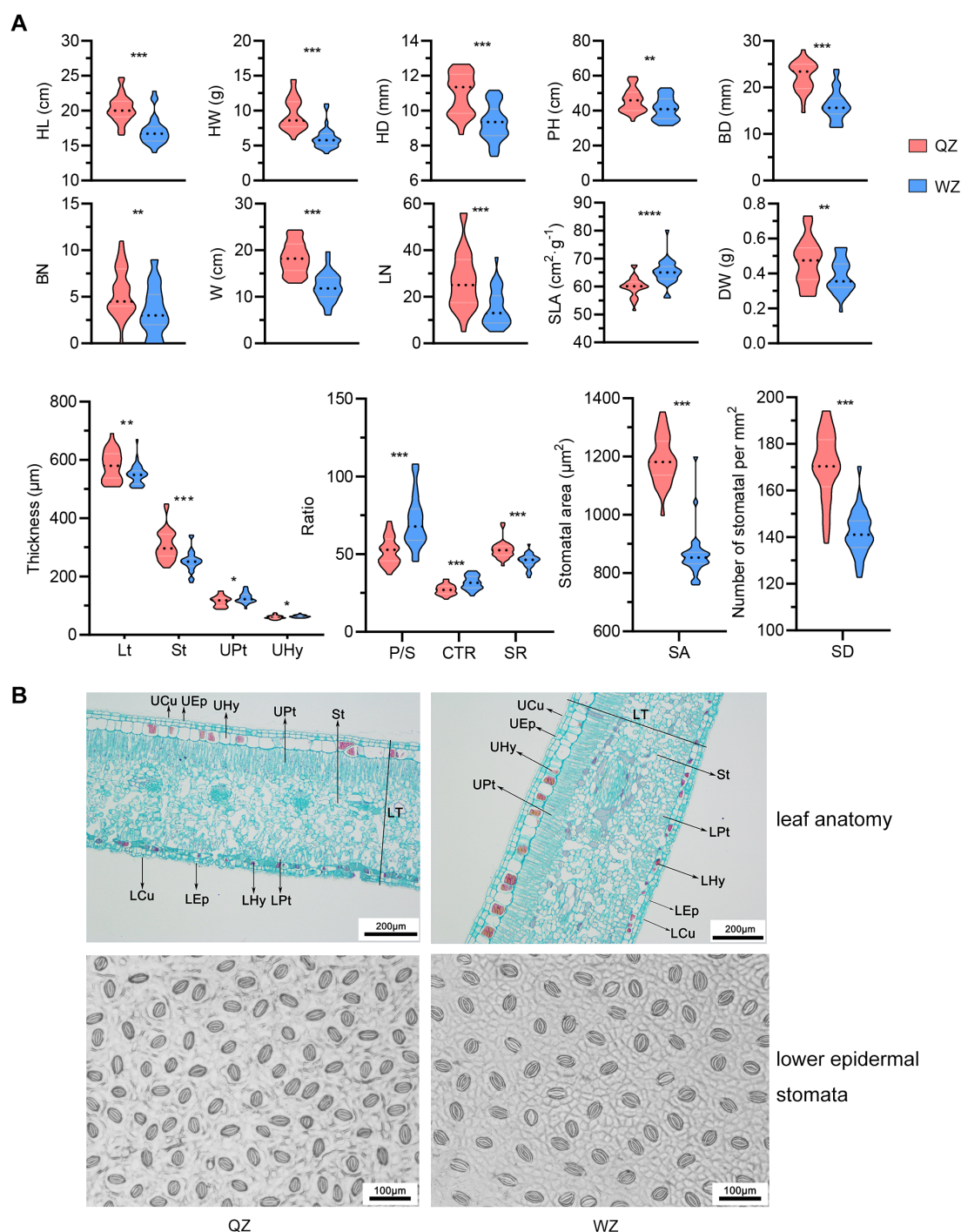


FIGURE 1

Differences in morphology and leaf traits between two populations of *K. obovata*. **(A)** Comparison of morphological traits between two populations of *K. obovata*. HL, hypocotyl length; HW, hypocotyl weight; HD, hypocotyl diameter; PH, plant height; BD, basal diameter; BN, branch number; W, crown width; LN, leaf number; SLA, specific leaf area; DW, dry weight; Lt, leaf thickness; St, spongy tissue thickness; UPt, upper palisade tissue; UHy, upper hypodermis; P/S, palisade-spongy tissue ratio; CTR, cell tense ratio; SR, spongy ratio; SA, stomatal area; SD, stomatal density. * represents $p < 0.05$; ** represents $p < 0.01$; *** p represents < 0.001 . **(B)** Comparison of leaf anatomy and lower epidermal stomata between two populations of *K. obovata*. UCu and LCu, upper and lower cuticles; UEp and LEp, upper and lower epidermis; UHy and LHy, upper and lower hypodermis; UPt and LPt, upper and lower palisade tissue; St, spongy tissue; LT, leaf thickness.

TABLE 1 Natural overwinter survival rates of *K. obovata* in 2020 and 2021.

Populations	numbers	No extreme low temperature (2020)	With extreme low temperature (2021)
WZ	19600	84.5%	18.6%
QZ	19400	67.3%	6.8%
Significant difference	–	$p < 2.2 \times 10^{-16}$	$p < 2.2 \times 10^{-16}$

Note that there are significant differences in the survival rates of the WZ and QZ populations in both 2020 and 2021.

3.4 WGRS, SNP variation detection and annotation

In this study, we selected samples from four population: DZG as the outgroup, ZZ as the sample from original location, QZ, and WZ. 10 samples were collected from each population, resulting in a total of 40 samples. WGRS was performed on these 40 samples. Following rigorous quality control and raw data filtering, we obtained a total of 92.98 Gb of clean read data with Q30 sequences exceeding 93.36% and GC content ranging from 36.35% to 39.64% (Supplementary Table 2). Alignment of the clean data with the reference genome yielded 45.82 Gb of clean data, indicating an average alignment rate of 92.62% between the samples and the reference genome. The genome coverage rate was calculated at 90.14%, with an average sequencing depth of 12.20× and an average coverage rate ($\geq 4\times$) (Supplementary Table 3). The uniform coverage across the entire genome indicated excellent sequencing randomness (Supplementary Figure 3).

After filtration, a total of 117,774 high-quality SNPs were identified in the population. Subsequent analysis revealed 69,692 transitions (ts) and 48,082 transversions (tv), resulting in a ts/tv ratio of 1.45 (Supplementary Table 4). Statistical analysis based on the annotation information of the *K. obovata* genome indicated that the majority of SNPs were located in intergenic regions, with only approximately 20% found within genic regions (Supplementary Table 4). Specifically, whole-genome assessment of WZ and QZ identified a total of 50,330 and 44,418 SNPs respectively, with approximately 20.68% and 19.40% located within genic regions (Supplementary Figure 4).

3.5 Analysis of genetic diversity and relationship among different populations of *K. obovata*

SNP analysis was employed to investigate the genetic relationships among 40 samples. The DZG population was utilized as an outgroup, and a phylogenetic tree was constructed using the maximum likelihood method. The results revealed challenges in distinguishing between the ZZ and QZ populations, indicating low genetic differentiation and close genetic relationship. However, the WZ population exhibited clear differentiation from

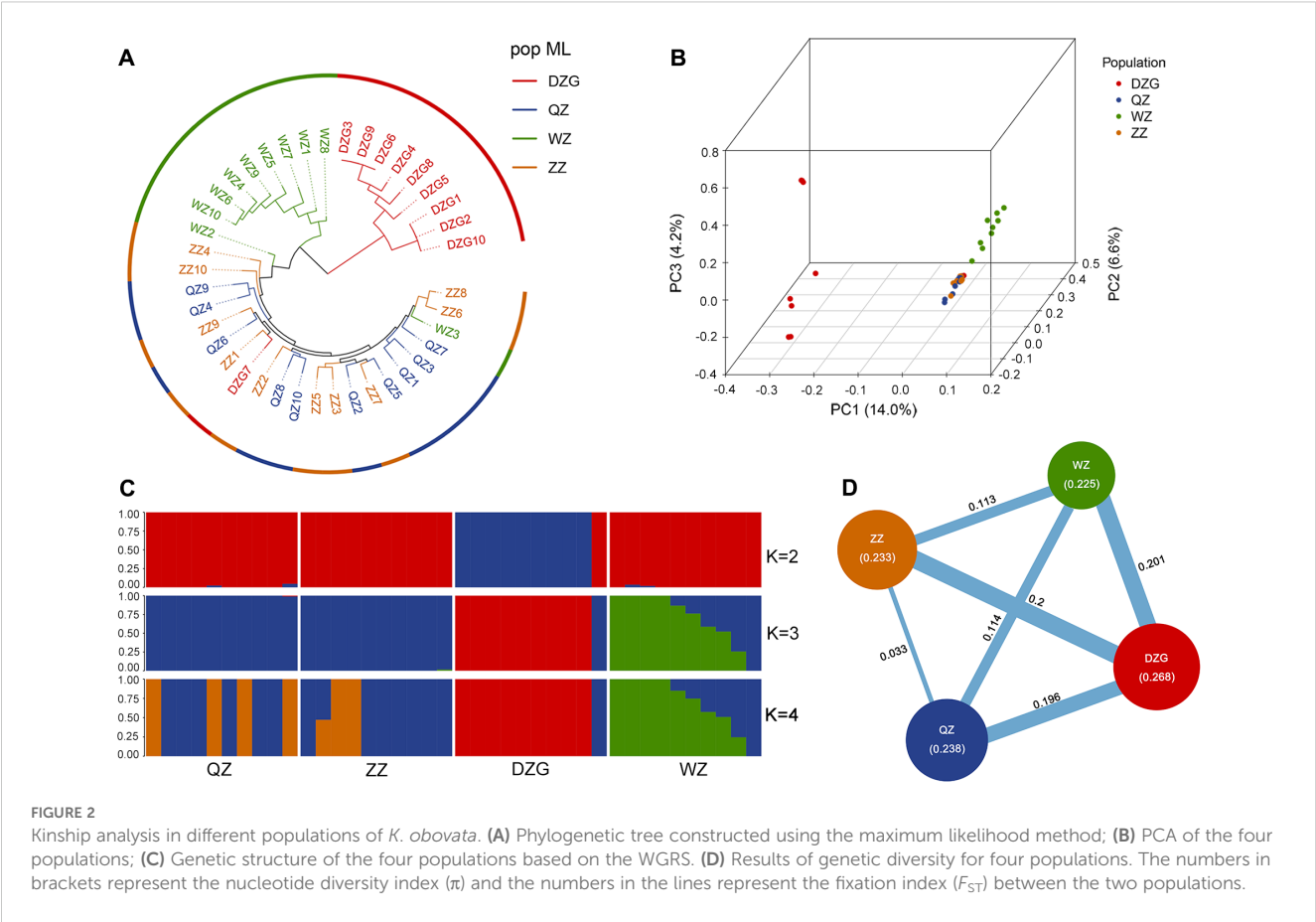
the ZZ and QZ populations, suggesting substantial genetic divergence between WZ and ZZ/QZ populations (Figure 2A). PCA revealed that PC1 effectively distinguished the DZG outgroup from other populations, while PC2 differentiated the WZ population from ZZ and QZ populations. Only a slight difference was observed between ZZ and QZ populations on this principal component, indicating their close genetic relationship (Figure 2B). STRUCTURE analysis revealed that a K value of 3 resulted in the lowest error rate of variation coefficient (Supplementary Figure 5), leading to the division of the 40 samples into three clusters. This grouping combined QZ and ZZ into one cluster while separating DZG and WZ (Figure 2C).

In addition, we computed π , Ho, He, PIC, and Shannon index for each population of *K. obovata* to evaluate their genetic diversity levels. The findings revealed that the DZG population exhibited the highest genetic diversity level among the four *K. obovata* populations, followed by QZ and ZZ; whereas the WZ population displayed the lowest genetic diversity level (Table 2). Furthermore, we assessed inter-population genetic differentiation using genetic differentiation index F_{ST} . The results indicated minimal genetic differentiation between the QZ and ZZ populations, while moderate genetic differentiation was observed between the WZ and QZ as well as ZZ populations (Figure 2D).

3.6 Detection of selection sweep signals and analysis of *cis*-elements in candidate genes

Using the 1,000 permutation test, we observed that the 95th percentile of the F_{ST} distribution was 0.243 in the absence of selection pressure. In contrast, the real population's top 5% F_{ST} threshold was 0.534 (Supplementary Figure 6). The fact that the observed F_{ST} value exceeds the 95th percentile of the permutation distribution indicates that these high F_{ST} values are not due to random fluctuations and are more likely to reflect genuine selection pressures. Thus, the top 5% regions of F_{ST} in *K. obovata* populations from QZ and WZ were identified based on F_{ST} scans, resulting in the detection of 785 windows through Manhattan plots. Following alignment with the reference genome annotation file, a total of 1,159 genes were identified (Figure 3A, Supplementary Tables 5–7). KEGG pathway enrichment analysis of these 1,159 candidate genes revealed significant enrichment in 9 pathways (Supplementary Table 8). Specifically, 15 genes showed enrichment in cysteine and methionine metabolism, while 9 genes exhibited enrichment in glutathione metabolism, along with 8 genes involved in alanine, aspartate, and glutamate metabolism; additionally, 7 genes were associated with ubiquinone and other terpenoid-quinone biosynthesis (Figure 3B).

Furthermore, employing a method that integrates genetic differentiation coefficient (F_{ST}) and nucleotide diversity (π), we identified 13 regions exhibiting positive selection in WZ, falling within the top 5% of the region between F_{ST} and $\log_2(\pi_{QZ}/\pi_{WZ})$ in WZ (Figure 3C). Upon aligning these selected windows with the reference genome annotation file, we determined a total of 40



candidate genes undergoing positive selection in WZ (Table 3; Supplementary Table 9). GO functional enrichment analysis revealed significant enrichment of biological processes related to stress response, anatomical structure development, sexual reproduction, reproductive development process, and post-embryonic development among the positively selected genes in WZ (Figure 3D, Supplementary Table 10).

We utilized the 2,000 bp upstream sequence as the promoter region for investigating 40 genes exhibiting positive selection in the WZ population. Our analysis revealed an abundance of *cis*-elements associated with non-biological stress responses (Supplementary Figure 7). The findings demonstrate widespread presence of stress-responsive elements (STR), low temperature-responsive elements (LTR), MYB-binding sites (MBS) induced by drought, and wound-responsive elements (WUN-motif) within the promoters of positively selected genes in the WZ population.

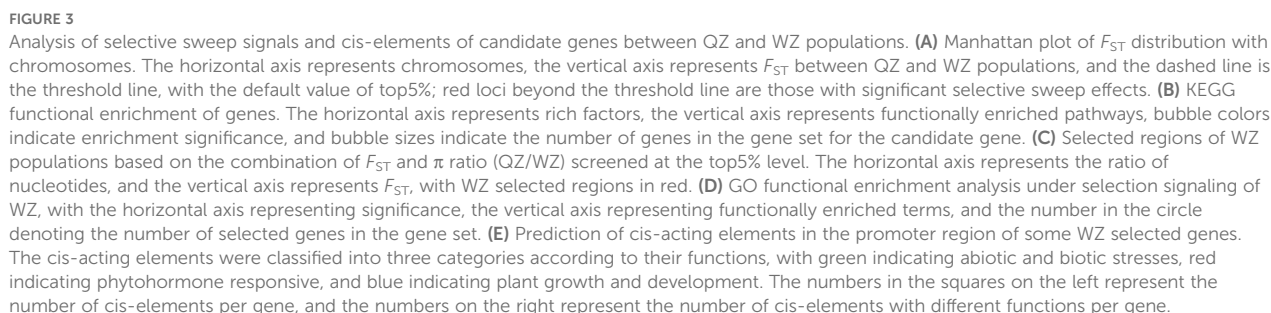
TABLE 2 Genetic diversity index statistics among *K. obovata* populations.

Pop ID	π	Ho	He	PIC	Shannon Index
DZG	0.268	0.242	0.252	0.199	0.374
QZ	0.238	0.257	0.223	0.177	0.334
ZZ	0.233	0.256	0.219	0.175	0.329
WZ	0.225	0.230	0.211	0.169	0.317

Furthermore, we have observed a wide distribution of diverse *cis*-elements, including abscisic acid responsive element (ABRE), salicylic acid responsive element (as-1), ethylene responsive element (ERE), jasmonic acid responsive elements (CGTCA-motif and TGACG-motif), as well as numerous light-responsive elements such as Box4, across these promoters. (Figure 3E).

3.7 Analysis of the relationship between selected genes and cold stress response in the WZ population

In our previous study, we examined the transcriptome profile of QZ and WZ during a simulated cold wave (Supplementary Figure 8), identifying 3,810 DEGs (Zhang et al., 2024). In this study, we conducted an analysis of 1,159 genes exhibiting positive selection, as identified from the F_{ST} index scan (within the top 5%) of the ZZ and WZ populations based on resequencing data. Upon taking the intersection of these gene sets, we obtained 215 genes (Figures 4A, B; Supplementary Table 11), which have shown differential transcription changes in the cold wave (Figure 4C). These genes were subsequently subjected to SNP mining and functional annotation analysis. Our findings revealed a total of 592 variants, including 206 variations within 1 kb region upstream of transcription start site (TSS), 195 variations within 1 kb region downstream of transcription end site, 96 variations within introns,



The differential expression of genes can be attributed to the presence of distinct *cis*-acting elements within various gene promoters. Variations in upstream sequences have the potential to modify these *cis*-acting elements and binding sites for transcription factors, thereby influencing their binding capacity and efficiency, consequently impacting gene expression levels. Genes exhibiting

variations in their upstream sequences were then screened, and promoter analysis was conducted for these varied upstream sequences. Substantial disparities in *cis*-regulatory elements between the WZ and QZ populations have been found. For instance, within geneMaker00002377, substitution of A for G at nucleotide position 1656 in the QZ population results in an additional stress response element (STRE) present exclusively in the WZ population. Similarly, within the geneMaker00002517 encoding drought-induced protein,

TABLE 3 Candidate genes identified by the top5% F_{ST} and π ratio value between WZ and QZ populations (40 genes).

Gene ID	Gene Symbol in Arabidopsis	Description
geneMaker00000121	<i>LRL12</i>	Leaf rust 10 disease-resistance locus receptor-like protein kinase-like 1.2
geneMaker00000483	<i>FH20</i>	LOW QUALITY PROTEIN: formin-like protein 20 [Populus alba]
geneMaker00000499	<i>ENO1</i>	enolase 1, chloroplastic [Manihot esculenta]
geneMaker00000704	<i>AAE17</i>	probable acyl-activating enzyme 17, peroxisomal isoform X1 [Populus alba]
geneMaker00000809	<i>CDPKA</i>	calcium-dependent protein kinase 10-like [Hevea brasiliensis]
geneMaker00002020	<i>EPFL2</i>	EPIDERMAL PATTERNING FACTOR-like protein 2 [Hevea brasiliensis]
geneMaker00002335	<i>TTL3</i>	Inactive TPR repeat-containing thioredoxin TTL3
geneMaker00003382	<i>NOV</i>	Protein NO VEIN
geneMaker00004151	<i>EMC1</i>	ER membrane protein complex subunit 1 [Hevea brasiliensis]
geneMaker00004198	<i>ELF32</i>	ELF3-like protein 2
geneMaker00004589	<i>KTN83</i>	Katanin p80 WD40 repeat-containing subunit B1 homolog KTN80.3
geneMaker00005300	<i>PILS6</i>	protein PIN-LIKES 6-like isoform X2 [Hevea brasiliensis]
geneMaker00005988	<i>ACA9</i>	calcium-transporting ATPase 9, plasma membrane-type isoform X1 [Manihot esculenta]
geneMaker00006095	<i>ILL3</i>	IAA-amino acid hydrolase ILR1-like 3 [Hevea brasiliensis]
geneMaker00006313	<i>VOZ1</i>	Transcription factor VOZ1
geneMaker00006341	<i>TBL38</i>	protein trichome birefringence-like 37 [Manihot esculenta]
geneMaker00007113	<i>PSL5</i>	probable glucan 1,3-alpha-glucosidase [Hevea brasiliensis]
geneMaker00007422	<i>DPNP1</i>	SAL1 phosphatase
geneMaker00007597	<i>FIGL1</i>	ATPase family AAA domain-containing protein FIGL1 isoform X1 [Manihot esculenta]
geneMaker00008043	-	uncharacterized protein LOC110637836 [Hevea brasiliensis]
geneMaker00008215	<i>PERK3</i>	probable receptor-like protein kinase At5g38990 isoform X1 [Hevea brasiliensis]

(Continued)

TABLE 3 Continued

Gene ID	Gene Symbol in Arabidopsis	Description
geneMaker00008285	-	uncharacterized protein LOC110637868 isoform X3 [Hevea brasiliensis]
geneMaker00009206	-	uncharacterized protein LOC110658574 [Hevea brasiliensis]
geneMaker00009300	<i>GAUT6</i>	probable galacturonosyltransferase 6 [Ricinus communis]
geneMaker00009877	-	hypothetical protein H0E87_011124 [Populus deltoides]
geneMaker00009926	-	PREDICTED: uncharacterized protein LOC105131445 [Populus euphratica]
geneMaker00009969	<i>TGH</i>	G patch domain-containing protein TGH [Hevea brasiliensis]
geneMaker00010131	<i>DHAR2</i>	Glutathione S-transferase DHAR2
geneMaker00010295	<i>CPP1</i>	protein CHAPERONE-LIKE PROTEIN OF POR1, chloroplastic [Jatropha curcas]
geneMaker00010602	<i>SPO11</i>	meiotic recombination protein SPO11-1 isoform X3 [Hevea brasiliensis]
geneMaker00011284	<i>TAZ</i>	tafazzin [Manihot esculenta]
geneMaker00012218	<i>SMU1</i>	suppressor of mec-8 and unc-52 protein homolog 1 [Morus notabilis]
geneMaker00012254	<i>SYKC</i>	lysine-tRNA ligase-like isoform X3 [Gossypium australe]
geneMaker00012303	-	hypothetical protein MANES_17G067700v8 [Manihot esculenta]
geneMaker00012523	<i>SUV3M</i>	DEXH-box ATP-dependent RNA helicase DEXH16, mitochondrial isoform X2 [Ricinus communis]
geneMaker00015184	-	uncharacterized protein LOC18055655 isoform X1 [Citrus clementina]
geneMaker00015878	<i>AP2A1</i>	AP-2 complex subunit alpha-1
geneMaker00015905	<i>LAC4</i>	laccase-4-like [Hevea brasiliensis]
geneMaker00015962	<i>DIVARICATA</i>	transcription factor DIVARICATA [Herrania umbratica]
geneMaker00016914	<i>SIGF</i>	RNA polymerase sigma factor sigF, chloroplastic isoform X1 [Manihot esculenta]

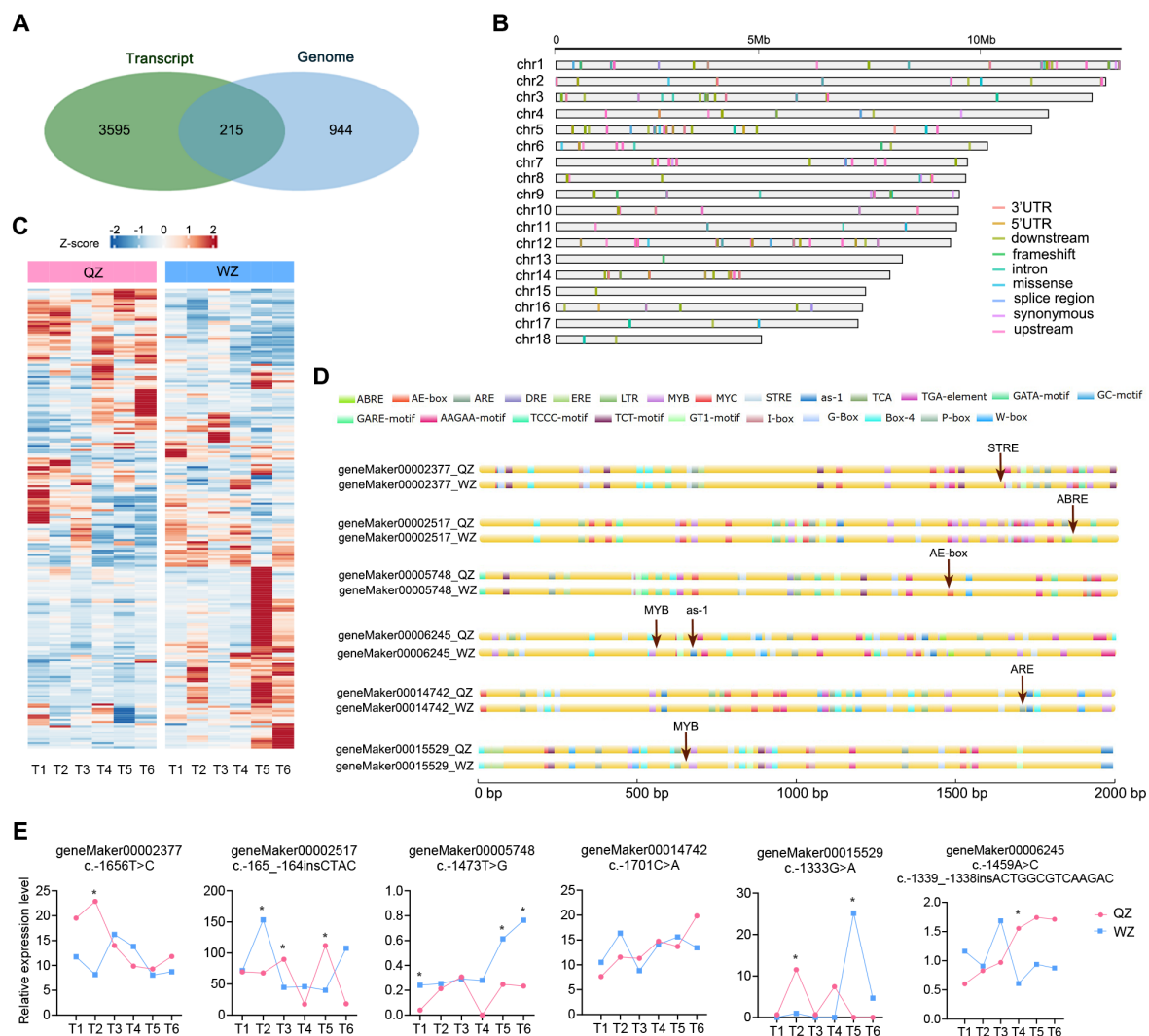


FIGURE 4

Integrated analysis of transcriptome and genome. (A) A Venn diagram illustrating the overlap of genes identified through F_{ST} -based genomic positive selection analysis and transcriptome DEGs under cold wave, resulting in 215 genes. (B) Visualization of SNP variation types and their chromosomal locations within the 215 genes, with each row representing a distinct chromosome and different colors denoting various variation types. (C) A heatmap displaying the transcription level changes of the 215 genes under cold wave. (D) The impact of upstream base mutations on *cis*-elements of *K. obovata* genes in two populations, indicated by red arrows and text depicting *cis*-elements before and after mutations. (E) The differential transcription levels of selected genes in the cold wave (* $p < 0.05$).

there is an extra abscisic acid response element (ABRE) present solely in the WZ population compared to QZ; this element can be recognized by bZIP transcription factors leading to increased expression under cold conditions and thus enhancing cold tolerance of *K. obovata* (Sun et al., 2022). Furthermore, there is an elevated C allele frequency at nucleotide position 1473 within geneMaker00005748 in WZ population which encodes *K. obovata* homologue of E3 ligase SHOOT GRAVITROPISM9 (SGR9) (Nakamura et al., 2011), resulting in an additional AE-box response element which may enhance cold tolerance. Additionally, the substitution from G to A at nucleotide position 1333 within extracellular signal-regulated kinase (ERK) activator CEP14-encoding gene (geneMaker00015529) (Delay et al., 2013) leads to an inclusion of MYB element exclusively found in WZ that could potentially bolster plant recovery ability under low temperature stress

during later stages (Figures 4D, E; Table 4). We propose that these differences may account for alterations in target gene transcription levels between the two populations and potentially contribute to changes in plant cold tolerance.

We also identified a significant number of mutations in the downstream gene sequences of DEGs. For example, the *CBF3* gene (geneMaker00008995), which constitutes the core of the cold stress pathway and plays an essential upstream initiation role in *K. obovata* cold resistance (Peng et al., 2020), exhibits an AA base deletion at locus 938050 on chromosome 9 in the WZ population. Furthermore, the *ABI5* gene (geneMaker00005324), a negative feedback factor in abscisic acid (ABA)-signaling involved in regulating ABA signaling pathway and ROS levels (Collin et al., 2021), shows a C to A base replacement at locus 2261919 on chromosome 5 within the WZ population. The potential

TABLE 4 Effects of upstream variation on *cis*-elements and functions of the *K. obovata*..

Gene id	Orthologs in Arabidopsis (Gene symbol)	Gene description	Mutation	Mutated cis-Acting Element	TF binding Sequence	TF	Function of Site
geneMaker00002377	–	–	c.-1656T>C	STRE	AGGGG	–	Cis-acting elements involved in stress responsiveness
geneMaker00002517	–	drought-induced protein	c.-165_-164insCTAC	ABRE	ACGTG	bZIP	cis-acting element involved in the abscisic acid responsiveness
geneMaker00005748	SHGR9	E3 ubiquitin-protein ligase SGR9, amyloplastic [Manihot esculenta]	c.-1473T>G	AE-box	AGAAACAA	Trihelix	part of a module for light response
geneMaker00014742	RIPK	probable serine/threonine-protein kinase PBL12 [Ricinus communis]	c.-1701C>A	ARE	AAACCA	MYB recognition site	cis-acting regulatory element essential for the anaerobic induction
geneMaker00015529	PCP14	Precursor of CEP14	c.-1333G>A	MYB	TAACCA	MYB recognition site	MYB binding site
geneMaker00006245	PTL	trihelix transcription factor PTL-like [Hevea brasiliensis]	c.-1459A>C	MYB	TAACCA	MYB recognition site	MYB binding site
			c.-1339_-1338insACTGGCGTCAAGAC	as-1	TGACG	bZIP	cis-acting element involved in salicylic acid responsiveness

relationship between these variants and changes in transcription levels of corresponding genes remains uncertain yet.

4 Discussion

4.1 Variation in phenotypic and leaf functional traits of *K. obovata* from different populations

Under the influence of natural and artificial selection, plants will inevitably differentiate in terms of phenotype and ecological traits as they adapt to new environments, gradually forming distinct geographical sources and leading to significant differences in morphology and functional leaf traits. Morphological indices such as leaf thickness, leaf dry matter content, leaf water content, and specific leaf area can reflect the plant's adaptive features and ability to acquire resources in different habitats. Specific leaf area is typically closely linked to the plant's growth strategies for survival. Plants with low specific leaf area often thrive in harsher or less fertile environments, while those with high specific leaf area can effectively maintain their nutrient content (Meziane and Shipley, 1999). Furthermore, as an index of ecological adaptation, leaf anatomical structure is closely related to a plant's cold tolerance; it is frequently used as a key metric for evaluating plant cold tolerance. Generally speaking, the higher the ratio of palisade to mesophyll tissue density within leaves (and thus tighter structural density), the greater their cold tolerance becomes. In response to adverse conditions such as high temperatures, small dense stomata are advantageous for avoiding rapid water loss due to transpiration.

QZ and WZ populations have been introduced into Fujian Quanzhou and Zhejiang Wenzhou within the last 20 years. Despite this study's elimination of native habitat effects through a CGE, these populations have already displayed distinctive phenotypic variations and leaf traits. The research findings suggest that the QZ population at lower latitudes exhibits greater PH, BD, W, LN and BN. Moreover, there are notable disparities in leaf functional traits and anatomical structures between the two populations. The WZ population demonstrates a larger SLA, potentially attributed to its robust photosynthetic capacity facilitating rapid resource acquisition and internal nutrient maintenance. Its palisade tissue is thicker with a denser cell structure, resulting in a higher P/S that enhances cold resistance. Conversely, the QZ population features smaller SA and higher SD possibly linked to relatively elevated local temperatures and strong transpiration for favorable water balance. In contrast, the WZ population displays larger but sparser stomata which may enhance its resilience to adverse conditions by reducing transpiration rate (Figures 1A, B).

It is widely accepted that there exists a trade-off between plant growth and defense. When plants are frequently exposed to pathogens or adverse environmental conditions, they prioritize defense mechanisms, leading to a reduction in growth and reproductive capacity (Zhou et al., 2022). We posit that the seedlings of WZ adopted a distinct energy allocation strategy

compared to those of QZ in order to more effectively acclimate to the high-latitude winter environment. Additionally, the functional characteristics of the leaves underwent changes aimed at enhancing their adaptability to the environment, with various leaf indices demonstrating a synergistic and trade-off relationship.

Consequently, we postulate that the WZ population of *K. obovata* inhabiting relatively cold environments may exhibit adaptive responses by modifying its growth-defense trade-off strategy. This could involve allocating more resources to defense mechanisms, slowing down the growth rate, increasing leaf tissue density and structural compactness, reducing SLA, and regulating stomatal characteristics as an adaptive response to the challenging habitat. These adaptive adjustments are likely to enhance the plant's resistance to low temperatures and improve its environmental adaptability and resource utilization. Conversely, the QZ population in milder habitats may allocate greater energy towards growth and reproduction. In conclusion, this study unveils significant phenotypic disparities between WZ and QZ, suggesting the adaptive capacity of *K. obovata* to environmental fluctuations. The heterogeneity of the environment over the approximately 20-year adaptation period may underlie the observed differentiation in phenotypic traits, leaf functionality, and physiological characteristics between QZ and WZ populations.

4.2 Genetic diversity and kinship analysis of different populations in the same introduction

The research conducted by Yang et al. demonstrates a substantial variation in both hypocotyl and seedling growth traits of *K. obovata* across diverse collection sites, showing a strong correlation with geographical and climatic factors (Yang et al., 2020). Prolonged acclimatization to its indigenous habitat can result in genetic divergence concerning plant morphology and leaf functional traits, thereby ensuring stable inheritance of these variations across successive generations. Our findings demonstrate that the mangrove species *K. obovata*, which was introduced from ZZ to WZ and QZ within a brief two-decade period, rapidly acclimated to the local environment through the development of specific traits. Significant phenotypic differences were already evident in the hypocotyls (Figure 1; Supplementary Figure 1).

The CGE further confirms that *K. obovata* seedlings also demonstrate significant variability in phenotype, leaf functional traits, and cold tolerance between the QZ and WZ populations (Figure 1, Table 1). The results of PCA and phylogenetic tree analysis based on WGRS indicate that ZZ and QZ are intertwined and clustered into the same branch, while WZ forms a separate cluster (Figures 2A, B). Additionally, there exists moderate genetic differentiation between the WZ and QZ populations, with experimental findings consistent with the geographic distribution of *K. obovata*. These results suggest that the observed phenotypic differences between the two *K. obovata* populations are primarily attributed to genetic differentiation rather than phenotypic plasticity.

This study utilized WGRS technology to assess the genetic diversity levels in various *K. obovata* populations. The analysis involved calculating heterozygosity (He, Ho), nucleotide diversity, polymorphic information content, and Shannon index. The findings revealed a decrease in genetic diversity of *K. obovata* populations in China with increasing latitude. At the whole-genome level, DZG exhibited the highest genetic diversity at lower latitudes, while ZZ and QZ displayed similar levels of genetic diversity and WZ showed the lowest (Table 2). Furthermore, observed heterozygosity (Ho) for ZZ, QZ, and WZ populations exceeded expected heterozygosity (He) (Table 2), suggesting higher frequency fluctuations possibly due to founder effect and bottleneck effect during the northward expansion of *K. obovata* driven by human selection of breeding individuals. The initial genetic bottleneck during the establishment of the WZ population likely resulted in limited genetic variation, leading to random phenotypic changes, some of which may be attributed to the founder effect. However, as the WZ population adapted to low-temperature environment, natural selection likely accelerated the optimization of these variants, driving the positive selection of cold-tolerance-related genes. For instance, we identified significant enrichment in pathways such as “related to stress response,” and “glutathione metabolism” (Figure 3D), along with selection signals linked to ABA/JA and low-temperature response elements, which are crucial for cold stress adaptation (Figure 3E). The WZ population has developed robust cold tolerance under prolonged exposure to low temperatures, and this trait has remained genetically stable across multiple generations, as confirmed by CGE in Shanghai (Table 1; Figure 1). The WZ population consistently exhibited enhanced overwinter survival rates (Table 1) and stronger cold-resistant phenotypes, accompanied by adaptive leaf anatomical changes, including thicker palisade tissue and a higher palisade-to-spongy ratio (Figure 1). These traits showed consistent inheritance, supporting the role of adaptive evolution in shaping this phenotype. Thus, while the founder effect may have initially constrained genetic diversity in the WZ population, the enhanced cold tolerance is more likely the result of adaptive evolution driven by natural selection. In conclusion, the phenotypic changes observed in the WZ population reflect a combination of adaptive evolution and the founder effect. Further studies, such as whole-genome association studies (GWAS) or quantitative trait locus (QTL) analysis, are needed to explore other phenotypic differences beyond cold tolerance.

Transcription factors regulate the expression of a wide range of biological and abiotic stress response genes by interacting with *cis*-acting elements in the promoter region, enabling plants to adapt to diverse stresses. To further explore potential stress response elements in the promoters of robust WZ populations, we utilized the Plantcare software for *cis*-element prediction in their promoter regions. The analysis revealed a substantial presence of *cis*-elements associated with abiotic stress response, such as LTR and hormone response elements (Figure 3E), suggesting their significant role in plant growth, development, low-temperature stress responses, and modulation of hormone signaling molecules. Endogenous plant hormones play a pivotal role in abiotic stress (Waadt et al., 2022).

This investigation identified ABRE within almost all positively selected genes. Furthermore, most genes contain jasmonic acid (JA) response elements CGTCA-motif and TGACG-motif within their promoter sequences which facilitate ABA- and JA-dependent signaling pathways through interaction with upstream transcription factors. ABA and JA, essential endogenous plant hormones, are crucial in mediating abiotic stress responses and have been shown to significantly impact on low temperature stress in plants (Huang et al., 2017; Hu et al., 2017). Studies have shown that exogenous ABA application can significantly improve cold tolerance in many plants, such as *Capsicum annuum* (Guo et al., 2012) and *Vitis vinifera* (Wang et al., 2020). Similarly, JA positively regulates the ICE-CBF pathway to enhance cold tolerance in *Arabidopsis thaliana*, and exogenous application of methyl jasmonate (MeJA) significantly improves cold tolerance in plants (Hu et al., 2013, Hu et al., 2017). Notably, exogenous ABA can alleviate cold stress in *K. obovata* by activating antioxidant enzyme activities and promoting the accumulation of osmotic regulators, thereby mitigating the negative effects of cold stress (Liu et al., 2022). Furthermore, in our previous research, we found that the enhanced cold tolerance in the WZ population is linked to JA signaling molecules and exogenous application of MeJA reduced cold-induced damage in the QZ population (Zhang et al., 2024). These findings suggest that these genes may be regulated by ABA and JA pathways thereby promoting *K. obovata*'s response to low temperature stress.

4.3 Low Temperature Maybe a Major Factor in the Genetic Differentiation of *K. obovata*

Climate exerts significant selective pressure on plant traits, with low temperature playing a crucial role in influencing the distribution of mangrove plants at high latitudes. Previous studies have shown that temperature is a major environmental factor limiting the distribution of mangroves (Cavanaugh et al., 2014). For instance, extreme freezing events are considered as a primary environmental factor determining the latitudinal distribution of *K. obovata* (Su et al., 2019; He et al., 2023) or other species such as *Avicennia germinans* (Pickens and Hester, 2011; Osland et al., 2014). A related study integrated geographical distribution data for *K. obovata* from both natural and introduced populations along China's southeastern coast and combined it with climate and hydrological data to quantitatively analyze the relationship between geographical distribution patterns and key environmental factors (Zhu et al., 2021). The results indicated that the top five climate factors influencing *K. obovata*'s distribution were annual mean temperature, mean temperature of the coldest quarter, extreme minimum temperature, temperature seasonality, and mean temperature of the driest quarter. These findings are consistent with our observations. On the one hand, phenotypic and genomic analyses revealed temperature-associated variations, including selection signatures at cold-responsive elements and

promoter polymorphisms in cold adaptation genes (Table 1; Figures 3, 4). On the other hand, based on data from the China Meteorological Science Data Sharing Service Network (<http://data.cma.cn>), the average temperature and lowest temperature of the coldest month in the WZ and QZ regions over the past 75 years are as follows: 8.1°C and 12.0°C for the average temperatures of the coldest month in the WZ and QZ regions respectively; -0.9°C and 5.5°C for their respective average extreme low temperatures in the coldest month (Supplementary Figure 9; Supplementary Table 13). These findings indicate that temperature likely serves as a primary driver of genetic differentiation between populations of *K. obovata* introduced from related source sites (WZ and QZ) due to substantial differences in latitude, average temperature, and other environmental conditions. However, it should be noted that while temperature gradients strongly correlate with latitudinal variation, other environmental parameters may synergistically shape genetic differentiation. Although WZ is located at a higher latitude and experiences colder temperatures, the observed genetic differentiation may result from a complex interplay of multiple environmental pressures, rather than temperature alone. Future studies that collect additional long-term environmental data—such as salinity, precipitation, and soil parameters—from both QZ and WZ will provide a more comprehensive analysis of the multi-layered environmental factors influencing the genetic structure and distribution of *K. obovata*.

Following transplantation to higher latitudes, acclimatization processes occurred within the ZZ population, leading to elimination of mangrove plants ill-suited for low temperatures while retaining those adapted to such conditions, ultimately giving rise to today's WZ population which demonstrates enhanced adaptability to low temperatures at high latitudes (Table 1). The molecular mechanisms underlying adaptation to low temperatures by the WZ population have been partially elucidated (Su et al., 2019). In summary, both natural selection and genetic drift play crucial roles in shaping the adaptation of *K. obovata* to its environment. After approximately 20 years, phenotypic and nucleotide-level genetic differentiation between the WZ and ZZ populations indicates that *K. obovata* possesses potential for rapid adaptation, rendering it a promising candidate species for studying plant adaptability.

5 Conclusion

In conclusion, this study investigated the phenotypic and functional trait variations of WZ and QZ populations in common garden experience. The findings revealed that different populations exhibited distinct growth-defense trade-off strategies to adapt to their respective environments. Within the same introduction site, moderate genetic differentiation was observed among *K. obovata* populations, indicating the rapid adaptation ability of *K. obovata* to environmental changes. WGRS further elucidated the genetic structure of *K. obovata* populations within the same introduction site, with results showing that the northernmost artificially introduced population (WZ) displayed the lowest genetic diversity. Additionally, its gene

promoter regions affected by strong selection contained a significant number of stress response elements related to low temperature and hormones, suggesting that temperature may be a primary driver of genetic differentiation in *K. obovata*. These research findings contribute to our understanding of environmental adaptation characteristics and growth/survival strategies of *K. obovata* in introduction areas, providing theoretical support for its cultivation as well as for mangrove forest protection and development efforts.

Data availability statement

The WGRS data presented in the study are deposited in the NCBI repository, accession number PRJNA1145277 (<https://www.ncbi.nlm.nih.gov/sra/PRJNA1145277>). The transcriptome data cited in this study are deposited in the NCBI repository, accession number PRJNA1093421 (<https://www.ncbi.nlm.nih.gov/sra/PRJNA1093421>).

Author contributions

JZ: Writing – original draft. SO: Writing – original draft. XC: Writing – original draft. SY: Writing – original draft. QC: Writing – original draft. JY: Writing – original draft. ZS: Writing – original draft. WZ: Writing – original draft. YW: Writing – original draft. YZ: Writing – original draft. PN: Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors declare that this study received funding from Morgan Stanley Securities (China) Co., Ltd. and Shanghai Genshi Medical Technology Development Co., Ltd. and the National Natural Science Foundation of China (32270629 and 32070201). The funders were not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

The present manuscript commemorates professor Yang Zhong's enduring vision and exceptional contributions to the mangroves in China.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.2025.1512620/full#supplementary-material>

References

- 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
- Bastias, C. C., Estarague, A., Vile, D., Gaignon, E., Lee, C. R., Exposito-Alonso, M., et al. (2024). Ecological trade-offs drive phenotypic and genetic differentiation of *Arabidopsis thaliana* in Europe. *Nat. Commun.* 15, 5185. doi: 10.1038/s41467-024-49267-0
- Baythavong, B. S., and Stanton, M. L. (2010). Characterizing selection on phenotypic plasticity in response to natural environmental heterogeneity. *Evolution* 64, 2904–2920. doi: 10.1111/j.1558-5646.2010.01057.x
- Bjorkman, A. D., Vellend, M., Frei, E. R., and Henry, G. H. (2017). Climate adaptation is not enough: warming does not facilitate success of southern tundra plant populations in the high Arctic. *Glob. Chang. Biol.* 23, 1540–1551. doi: 10.1111/gcb.13417
- Boquete, M. T., Muyle, A., and Alonso, C. (2021). Plant epigenetics: phenotypic and functional diversity beyond the DNA sequence. *Am. J. Bot.* 108, 553–558. doi: 10.1002/ajb2.1645
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354
- Cavanaugh, K. C., Kellner, J. R., Forde, A. J., Gruner, D. S., Parker, J. D., Rodriguez, W., et al. (2014). Poleward expansion of mangroves is a threshold response to decreased frequency of extreme cold events. *Proc. Natl. Acad. Sci. U.S.A.* 111, 723–727. doi: 10.1073/pnas.1315800111
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009
- Chen, L., Wang, W., Li, Q. Q., Zhang, Y., Yang, S., Osland, M. J., et al. (2017). Mangrove species' responses to winter air temperature extremes in China. *Ecosphere* 8, 10.1002/ecs2.1865
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi: 10.1093/bioinformatics/bty560
- Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)* 6, 80–92. doi: 10.4161/fly.19695
- Collin, A., Daszkowska-Golec, A., and Szarejko, I. (2021). Updates on the role of abscisic acid insensitive 5 (abi5) and abscisic acid-responsive element binding factors (ABFs) in ABA signaling in different developmental stages in plants. *Cells* 10, 10.3390/cells10081996
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330
- Delay, C., Imin, N., and Djordjevic, M. A. (2013). CEP genes regulate root and shoot development in response to environmental cues and are specific to seed plants. *J. Exp. Bot.* 64, 5383–5394. doi: 10.1093/jxb/ert332
- Ellison, A. M., Felson, A. J., and Friess, D. A. (2020). Mangrove rehabilitation and restoration as experimental adaptive management. *Front. Mar. Sci.* 7, 10.3389/fmars.2020.00327
- Foyer, C. H., and Kranner, I. (2023). Plant adaptation to climate change. *Biochem. J.* 480, 1865–1869. doi: 10.1042/bj20220580
- Freed, D., Aldana, R., Weber, J., and Edwards, J. (2017). The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. doi: 10.1101/115717
- Guo, W. L., Chen, R. G., Gong, Z. H., Yin, Y. X., Ahmed, S. S., and He, Y. M. (2012). Exogenous abscisic acid increases antioxidant enzymes and related gene expression in pepper (*Capsicum annuum*) leaves subjected to chilling stress. *Genet. Mol. Res.* 11, 4063–4080. doi: 10.4238/2012.September.10.5
- He, S., Wang, X., Du, Z., Liang, P., Zhong, Y., Wang, L., et al. (2023). Physiological and transcriptomic responses to cold waves of the most cold-tolerant mangrove, *Kandelia obovata*. *Front. Plant Sci.* 14, 10.3389/fpls.2023.1069055
- Hu, M. J., Sun, W. H., Tsai, W. C., Xiang, S., Lai, X. K., Chen, D. Q., et al. (2020). Chromosome-scale assembly of the *Kandelia obovata* genome. *Hortic. Res.* 7, 75. doi: 10.1038/s41438-020-0300-x
- Hu, Y., Jiang, L., Wang, F., and Yu, D. (2013). Jasmonate regulates the inducer of cbf expression-C-repeat binding factor/DRE binding factor1 cascade and freezing tolerance in *Arabidopsis*. *Plant Cell* 25, 2907–2924. doi: 10.1105/tpc.113.112631
- Hu, Y., Jiang, Y., Han, X., Wang, H., Pan, J., and Yu, D. (2017). Jasmonate regulates leaf senescence and tolerance to cold stress: crosstalk with other phytohormones. *J. Exp. Bot.* 68, 1361–1369. doi: 10.1093/jxb/erx004
- Huang, X., Shi, H., Hu, Z., Liu, A., Amombo, E., Chen, L., et al. (2017). ABA is involved in regulation of cold stress response in *Bermudagrass*. *Front. Plant Sci.* 8, 10.3389/fpls.2017.01613
- Jung, Y., and Han, D. (2022). BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* 38, 2404–2413. doi: 10.1093/bioinformatics/btac137
- Korunes, K. L., and Samuk, K. (2021). pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* 21, 1359–1368. doi: 10.1111/1755-0998.13326
- Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi: 10.1093/nar/30.1.325
- 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, C., Wang, F., Yang, P., Wang, F. C., Hu, Y. Z., Zhao, Y. L., et al. (2024). Mangrove wetlands distribution status identification, changing trend analysis and carbon storage assessment of China. *China Geology* 7, 1–11. doi: 10.31035/cg2023049
- Liu, X., Lu, X., Yang, S., Liu, Y., Wang, W., Wei, X., et al. (2022). Role of exogenous abscisic acid in freezing tolerance of mangrove *Kandelia obovata* under natural frost condition at near 32°N. *BMC Plant Biol.* 22, 593. doi: 10.1186/s12870-022-03990-2
- Lu, W. X., Zhang, B. H., and Yang, S. C. (2023). Survive the north: transplantation for conservation of mangrove forests requires consideration of influences of low temperature, mating system and their joint effects on effective size of the reforested populations. *Front. Ecol. Evol.* 11, 10.3389/fevo.2023.1160468
- Lu, W. X., Zhang, B. H., Zhang, Y. Y., and Yang, S. C. (2021). Differentiation of cold tolerance in an artificial population of a mangrove species, *Kandelia obovata*, is associated with geographic origins. *Front. Plant Sci.* 12, 10.3389/fpls.2021.695746
- Marchini, G. L., Maraist, C. A., and Cruzan, M. B. (2019). Trait divergence, not plasticity, determines the success of a newly invasive plant. *Ann. Bot.* 123, 667–679. doi: 10.1093/aob/mcy200
- 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
- Meziane, D., and Shipley, B. (1999). Interacting determinants of specific leaf area in 22 herbaceous species: effects of irradiance and nutrient availability. *Plant Cell Environ.* 22, 447–459. doi: 10.1046/j.1365-3040.1999.00423.x
- Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

- Miryeganeh, M. (2022). Mangrove forests: natural laboratories for studying epigenetic and climate changes. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.851518
- Nakamura, M., Toyota, M., Tasaka, M., and Morita, M. T. (2011). An Arabidopsis E3 ligase, SHOOT GRAVITROPISM9, modulates the interaction between statoliths and F-actin in gravity sensing. *Plant Cell* 23, 1830–1848. doi: 10.1105/tpc.110.079442
- Osland, M. J., Day, R. H., Larriviere, J. C., and From, A. S. (2014). Aboveground allometric models for freeze-affected black mangroves (*Avicennia germinans*): equations for a climate sensitive mangrove-marsh ecotone. *PLoS One* 9, e99604. doi: 10.1371/journal.pone.0099604
- Pathoumthong, P., Zhang, Z., Roy, S. J., and El Habti, A. (2023). Rapid non-destructive method to phenotype stomatal traits. *Plant Methods* 19, 36. doi: 10.1186/s13007-023-01016-y
- Peakall, R., and Smouse, P. E. (2012). GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research—an update. *Bioinformatics* 28, 2537–2539. doi: 10.1093/bioinformatics/bts460
- Peng, Y. L., Wang, Y. S., Fei, J., and Sun, C. C. (2020). Isolation and expression analysis of two novel C-repeat binding factor (CBF) genes involved in plant growth and abiotic stress response in mangrove *Kandelia obovata*. *Ecotoxicology* 29, 718–725. doi: 10.1007/s10646-020-02219-y
- Pickens, C. N., and Hester, M. W. (2011). Temperature tolerance of early life history stages of black mangrove *avicennia germinans*: implications for range expansion. *Estuaries Coasts* 34, 824–830. doi: 10.1007/s12237-010-9358-2
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675. doi: 10.1038/nmeth.2089
- Su, W., Ye, C., Zhang, Y., Hao, S., and Li, Q. Q. (2019). Identification of putative key genes for coastal environments and cold adaptation in mangrove *Kandelia obovata* through transcriptome analysis. *Sci. Total Environ.* 681, 191–201. doi: 10.1016/j.scitotenv.2019.05.127
- Sun, M. M., Liu, X., Huang, X. J., Yang, J. J., Qin, P. T., Zhou, H., et al. (2022). Genome-wide identification and expression analysis of the NAC gene family in *Kandelia obovata*, a typical mangrove plant. *Curr. Issues Mol. Biol.* 44, 5622–5637. doi: 10.3390/cimb44110381
- Waadt, R., Seller, C. A., Hsu, P. K., Takahashi, Y., Munemasa, S., and Schroeder, J. I. (2022). Plant hormone regulation of abiotic stress responses. *Nat. Rev. Mol. Cell Biol.* 23, 680–694. doi: 10.1038/s41580-022-00479-6
- Wang, H., Blakeslee, J. J., Jones, M. L., Chapin, L. J., and Dami, I. E. (2020). Exogenous abscisic acid enhances physiological, metabolic, and transcriptional cold acclimation responses in greenhouse-grown grapevines. *Plant Sci.* 293, 110437. doi: 10.1016/j.plantsci.2020.110437
- Wright, S. (1978). Evolution and the genetic of population, variability within and among natural populations. *Chicago: Univ. Chicago Press* 4, 213–220.
- Yang, S., Liu, X., Deng, R.-J., Chen, Q.-X., Wang, J.-W., and Lu, X. (2020). Geographic variations of hypocotyl and seedling growth traits for *Kandelia obovata* with different provenances. *Chin. J. Ecol.* 39, 1769–1777. doi: 10.13292/j.1000-4890.202006.003
- Zhang, J., Fan, T., Cai, X., Ouyang, S., Yang, J., Song, Z., et al. (2024). Comparative transcriptomic and metabolomic analysis of two related *Kandelia obovata* populations in response to cold wave. *Phenomics accepted*. doi: 10.1007/s43657-024-00204-7
- Zhao, C. P., Jia, M. M., Zhang, R., Wang, Z. M., Mao, D. H., Zhong, C. R., et al. (2024). Distribution of mangrove species *Kandelia obovata* in China using time-series sentinel-2 imagery for sustainable mangrove management. *J. Remote Sens.* 4, 15. doi: 10.34133/remotesensing.0143
- Zhao, Y., Zhong, Y., Ye, C., Liang, P., Pan, X., Zhang, Y. Y., et al. (2021). Multi-omics analyses on *Kandelia obovata* reveal its response to transplanting and genetic differentiation among populations. *BMC Plant Biol.* 21, 341. doi: 10.1186/s12870-021-03123-1
- Zhou, H., Hua, J., Zhang, J., and Luo, S. (2022). Negative interactions balance growth and defense in plants confronted with herbivores or pathogens. *J. Agric. Food Chem.* 70, 12723–12732. doi: 10.1021/acs.jafc.2c04218
- Zhu, H., Lin, H. J., Le, Y., Li, H. P., Yue, C. L., and Jiang, B. (2021). Geographical distribution pattern and environmental explanation of *Kandelia obovata* Sheue, H.Y. Liu & J. Yong populations along the Southeast coast of China. *Plant Sci. J.* 39, 476–487. doi: 10.11913/PSJ.2095-0837.2021.50476



OPEN ACCESS

EDITED BY

Robert Henry,
The University of Queensland, Australia

REVIEWED BY

Dejun Li,
Chinese Academy of Tropical Agricultural
Sciences, China
Alex Zaccaron,
Oregon State University, United States

*CORRESPONDENCE

Hanggui Lai

✉ laihanggui8938@163.com;

Dongyi Huang

✉ hdongyi@hainanu.edu.cn

[†]These authors have contributed equally to
this work

RECEIVED 06 January 2025

ACCEPTED 27 May 2025

PUBLISHED 30 June 2025

CORRECTED 23 July 2025

CITATION

Zhang S, Zhang Y, Luo S, Gao J, Hu H, Liu J,
Wu W, Wang J, Huang X, Lai H and Huang D
(2025) Mitochondrial genome assembly
and comparative analysis of
decaploid *Camellia hainanica*.
Front. Plant Sci. 16:1556379.
doi: 10.3389/fpls.2025.1556379

COPYRIGHT

© 2025 Zhang, Zhang, Luo, Gao, Hu, Liu, Wu,
Wang, Huang, Lai and Huang. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication
in this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Mitochondrial genome assembly and comparative analysis of decaploid *Camellia hainanica*

Shihui Zhang^{1†}, Yuyan Zhang^{1†}, Sheng Luo¹, Jie Gao¹,
Haiyan Hu¹, Jinping Liu¹, Wenqiang Wu², Jian Wang¹,
Xiaolong Huang², Hanggui Lai^{1,3*} and Dongyi Huang^{3*}

¹School of Tropical Agriculture and Forestry, Hainan University, Danzhou, China, ²School of Life and Health Sciences, Hainan University, Haikou, China, ³School of Breeding and Multiplication (Sanya Institute of Breeding and Multiplication), Hainan University, Sanya, China

Introduction: Decaploid *Camellia hainanica* is a new tea oil *Camellia* species discovered in recent years that is unique to Hainan. This species has high nutritional and medicinal value and shows strong adaptability in the growth process. Mitochondria play an important role in plant cells and have an independent genetic system. Therefore, assembling and annotating the mitochondrial genome function of decaploid *C. hainanica* is of great significance.

Methods: This study successfully assembled the mitochondrial genome of decaploid *C. hainanica* and comprehensively annotated its functional genes using the Nanopore sequencing platform.

Results: Results showed that the mitochondrial genome is 902,617 bp in length, with a typical circular structure and a guanine–cytosine content of 45.79%. The genome encodes 64 protein-coding genes and contains a total of 76 genes, including 40 mRNA, 32 tRNA, 3 rRNA, and 1 pseudogene. Tetranucleotide repeats accounted for 38.60% of the simple sequence repeats. Only two genes, *atp6* and *sdh4*, had a Ka/Ks ratio <1, whereas the Pi value of the *sdh3* gene had a maximum of 0.00374 in these regions, suggesting that the *sdh3* gene can be used as a molecular marker for the analysis of the mitochondrial genome of *C. hainanica*. From the relative synonymous codon usage (RSCU) analysis, 29 codons had RSCU values >1, 27 of which (93%) ended in A or U, indicating a bias for A/U endings is present in *C. hainanica*. During RNA editing, 48.24% (260 loci) of amino acids were changed from hydrophilic to hydrophobic, resulting in an increase in the hydrophobicity of the protein. Comparative analysis identified 34 homologous fragments between the mitochondrial and chloroplast genomes, with the longest fragment being 9,572 bp in length. Phylogenetic analysis of the genomes showed that the Hainanese and Vietnamese varieties of tea oil *Camellia* are sister species.

Discussion: Results confirmed that the mitochondrial genomes of Hainanese and Vietnamese tea oil *Camellia* underwent gene rearrangement. Results also provided key data support for the utilization and conservation of tea oil germplasm resources and the breeding of varieties and are of great significance for promoting genetic evolution research, genetic breeding, and identification of tea oil *Camellia*.

KEYWORDS

decaploid *Camellia hainanica*, mitochondrial genome, comparative analysis, RNA editing, analysis system development

1 Introduction

Tea oil *Camellia* (*Camellia oleifera* Abel.) is a small evergreen tree or shrub with a high seed oil content and economic cultivation value. Tea oil *Camellia* originates in China and has a long history of cultivation in Hainan, where it is one of the island's traditional plants (Liang et al., 2024).

Recent advances in *Camellia oleifera* genomics have revealed significant progress in understanding its genetic architecture and agronomic traits. While previous studies predominantly focused on diploid assemblies (Ye et al., 2023), emerging research highlights the imperative to investigate polyploid genomes. Current findings demonstrate that wild *C. oleifera* exists as a heterozygous hexaploid species characterized by elevated genomic heterozygosity (0.82%) and substantial repetitive element content (~68% of genome), as evidenced by Haoxing Xie (Xie et al., 2024) through PacBio HiFi sequencing. This breakthrough enabled the identification of 21,437 SSR markers and the characterization of cold-responsive genes including CBF and ICE1 transcription factors.

Notably, our investigation of Hainan Island germplasm uncovered a unique ploidy distribution pattern dominated by decaploid ($2n=10x=150$) and octaploid ($2n=8x=120$) cytotypes. Phylogenetic analyses suggest these polyploid complexes likely originated through adaptive radiation under the tropical monsoon climate regime, with the decaploid form representing an endemic lineage showing distinct evolutionary divergence ($FST > 0.25$). The genomic complexity of these cytotypes, particularly the decaploid's 15 Gb genome featuring 12.3% tandem repeats, presents both challenges and opportunities for resolving paleopolyploidization events through third-generation sequencing approaches.

These findings substantially expand the genomic resources for *Camellia* species while providing molecular tools for marker-assisted selection. The Hainan decaploid accessions, with their exceptional environmental adaptation mechanisms, hold significant promise for elucidating polyploid genome evolution and developing stress-resilient cultivars. Mitochondria harbor their own genetic code and protein translation system. Their DNA, known as mtDNA, encodes key components of cellular energy supply and participates in essential biological processes.

Unlike the plant chloroplast genome, the mitochondrial genome exhibits remarkable diversity due to lineage-specific evolutionary development (Wideman et al., 2020; Butenko et al., 2024). This diversity enables mitochondria to play crucial roles in energy conversion, fatty acid synthesis, amino acid metabolism, and stress responses, thereby enhancing their adaptability to changing environmental conditions and contributing to the adaptive evolution of plants. Although mtDNA is typically described as a circular molecule, diverse structures have been identified, including linear conformations, branching structures, and numerous smaller circular molecules (Sloan, 2013; Gualberto et al., 2014). These complex and varied structures harbor a vast amount of genetic information, which is invaluable for resolving species classification, accurate identification, and elucidating the evolutionary trajectories of species.

Most previous studies on *C. hainanica* focus on leaf characteristics and pollen spore analysis. However, the morphological characteristics of the plant are greatly affected by environmental changes. With advances in molecular biotechnology, many researchers have employed molecular biomarker technology and DNA sequencing methods to study species classification. The first complete assembly of the mitochondrial genome of the diploid tea oil species *C. gigantocarpa* was completed in 2022 (Lu et al., 2022). Using the PacBioHi-Fi and Hi-C sequencing technologies, the mitochondrial genome of *C. gigantocarpa* was successfully assembled. The proportion of repetitive sequences in the *C. gigantocarpa* mitochondrial genome is as high as 20.81%, comparable to that of *C. sinensis* (22.15%), but much higher than that of *Arabidopsis thaliana* (4.96%). This significantly increases the size of the mitochondrial genome of tea oil *Camellia*. In their analysis of the hexaploid tea oil variety *C. oleifera* cv. *Huashuo*, researchers have successfully assembled the full mitochondrial genome of this tea oil *Camellia* variety for the first time using second-generation sequencing technology. The study revealed a tea oil *Camellia* mitochondrial genome with a circular structure consisting of 709,596 bp and successfully annotated 74 genes in this genome (Gu et al., 2024). In a study of the mitochondrial genome of congener *C. assamica*, Rawal et al. (2020) obtained the complete *C. assamica* mitochondrial genome by de-redundancy

assembly of 587,142 filtered mitochondrial read sequences, obtaining a total length of 707,441 bp. The overall guanine–cytosine (GC) content was 45.75%, with a total of 66 annotated genes, including 35 protein-coding genes (PCGs), 29 tRNA, and 2 rRNA (Rawal et al., 2020). Li et al. (2024) studied the complete assembly and annotation of the mitochondrial genomes of 4 species within the tea oil *Camellia* lineage, demonstrating that the mitochondrial genome consists of closed-loop DNA molecules ranging in size from 850,836 bp (*C. nitidissima*) to 109,8121 bp (*C. tianeensis*) (Li et al., 2024). In a study on *C. duntosa*, Li et al. (2023b) demonstrated that its mitochondrial genome consists of 1,081,996 bp and 81 genes, including 1 pseudogene, 3 rRNA genes, 30 tRNA genes, and 47 PCGs (Li et al., 2023b). Although studies on the mitochondrial genomes of hexaploid tea oil *Camellia* and related species have been reported, reports regarding the mitochondrial genomes of higher ploidy tea oil *Camellia* varieties are currently lacking.

Based on this in-depth study of its mitochondrial genome, we aim to provide a scientific basis for the breeding and genetic improvement of tea oil *Camellia*. At the same time, comparative genomic analysis of decaploid *C. hainanica* with similar species can not only reveal the genetic differences between decaploid *C. hainanica* and other tea oil *Camellia* varieties but can also provide important data-based support for the utilization and protection of tea oil *Camellia* genetic resources, which can help promote the study of genetic evolution of the species and scientific identification in tea oil *Camellia* breeding.

2 Materials and methods

The *C. hainanica* in this study was sampled from Huishan Township, Qionghai City, Hainan Province (longitude 110°18'20" E, latitude 19°5'18"N, elevation 82.00 m), with the crops having been cultivated under natural conditions. The young leaves of perennial decaploid *C. hainanica* were collected, placed into liquid nitrogen for snap freezing, and stored in a –80°C freezer before being sent to Benagen for sequencing. The Plant Genomic DNA kit DP305 (Tiangen Biotech, Beijing, China) was used in this study. DNA purity was measured using a 1.0% agarose gel. To obtain an accurate full-length mitochondrial genome, short- and long-read sequencing technologies were combined in this study. The short-read sequencing platform used was Illumina NovaSeq 6000 (Illumina, San Diego, CA, USA), with a paired-end sequencing read length of 150 bp. The fastp (version 0.20.0; <https://github.com/OpenGene/fastp>) software was used to filter raw data and obtain high-quality reads. The long-read sequencing platform used was Nanopore PromethION (Nanopore, Oxford, UK), and the sequencing data were filtered by the filtlong software (version 0.2.1; <https://link.zhihu.com/?target=https%3A//github.com/rwwick/Filtlong>).

2.1 Mitochondrial genome assembly and annotation

Plant mitochondrial genes (coding sequence [CDS], rRNA) are highly conserved. By employing the third-generation alignment software minimap2 (Li, 2018), this characteristic was used to compare the third-generation data to the reference gene sequences (plant mitochondrial core genes, https://github.com/xul962464/plant_mt_ref_gene) and screened for sequences with an alignment length of >50 bp as candidate sequences for comparison. From these sequences, those with a larger number of aligned genes (sequences containing multiple core genes) and a higher alignment quality (the core genes covered were more complete) were selected as seed sequences. Next, minimap2 was used to align the original third-generation sequencing data to the seed sequences and screen for sequences with an overlap of >1 kb, which were then added to the seed sequences. Iterative alignment of the original data to the seed sequences was conducted, thus obtaining the complete third-generation sequencing data of the mitochondrial genome. Then, the third-generation assembly software canu (Koren et al., 2017) was used to correct the resulting third-generation data, and the corrected third-generation data were spliced using the default parameters of SPAdes (version 3.15.4, <https://github.com/ablab/spades#metapv>). The splicing results were visualized and manually adjusted using the Bandage (version 0.8.1) (Wick et al., 2015; <https://github.com/rwwick/Bandage>) software. Due to the complex physical structure of the mitochondrial genome that consists of multiple subloops, or even nonloops, the corrected third-generation sequencing data were aligned to the contig obtained from SPAdes (Andrey et al., 2020) using minimap2 to manually determine the branching direction, thereby obtaining the final assembly results.

Mitochondrial annotation was performed using the following steps:

1. Encoded proteins and rRNA were aligned to published and ref plant mitochondrial sequences using BLAST, with further manual adjustments made for closely related species.
2. tRNA was annotated using tRNAscanSE (5) (<http://lowelab.ucsc.edu/tRNAscan-SE/>).
3. Open reading frames were annotated using the Open Reading Frame Finder (<http://www.ncbi.nlm.nih.gov/gorf/gorf.html>) by setting the minimum length to 102 bp to exclude redundant sequences and sequences that overlap with known genes. Sequences >300 bp in length were annotated against the nr library.
4. RNA editing sites were originally predicted using PmtREP (<http://112.86.217.82:9919/#/tool/alltool/detail/336>). The final annotation results were obtained after checking and manually correcting the obtained results.

2.2 Synonymous codon usage bias analysis

The mitochondrial genome codon composition of *C. hainanica* was screened for unique CDS and calculated using a script written in Perl (<http://cloud.genepioneer.com:9929/#/tool/alltool/detail/214>). Its calculation method is: (the number of a certain codon encoding an amino acid/the number of all codons encoding that amino acid)/(1/the number of codon types encoding that amino acid), that is, (the actual usage frequency of the codon/the theoretical usage frequency of the codon).

2.3 Identification of RNA editing sites

The RNA sequencing data were aligned to the CDS (Coding DNA Sequence) sequences by utilizing Bowtie2 (version 2.3.5.1; <https://github.com/BenLangmead/bowtie2>) (Langmead and Salzberg, 2012), and subsequently processed using samtools (<https://github.com/samtools>) for further analysis. The software bcftools (1.9-170) (<https://github.com/samtools/bcftools>) was then used to identify sites where single-nucleotide polymorphisms existed between the sequencing data and the genome, which served as potential RNA editing sites.

2.4 Repeated sequence analysis

Repeated sequences include simple sequence repeats (SSRs), tandem repeats, and dispersed repeats. SSRs were identified using the misa software (version 1.0, parameters: 1-10 2-5 3-4 4-3 5-3 6-3, <https://webblast.ipk-gatersleben.de/misa/>), tandem repeats were identified using the trf software (trf409, parameters: 2 7 7 80 10 50 2000 -f -d -m, <http://tandem.bu.edu/trf/trf.submit.options.html>), and dispersed repeats were identified using BLASTn software (version 2.10.1, parameters: -word_size 7, E-value 1e-5, de-redundancy, tandem duplicates were removed) and visualized using circos v0.69-5.

2.5 Ka/Ks and Pi analyses

Binary grouping of the higher-order analyzed species was conducted to perform Ka/Ks analysis. Homologous gene pairs were then extracted, and the homologous gene pairs were aligned using the mafft version 7.427 (<https://mafft.cbrc.jp/alignment/software/>) software. After alignment, the KaKs_Calculator version 2.0 (Zhang, 2022) software was used to calculate the Ka and Ks values of each gene pair (<https://sourceforge.net/projects/kakscalculator2/>), with the MLWL calculation method selected.

Global alignment of homologous gene sequences from different species was performed using the mafft software (version 7.427, -automode), and the Pi values for each gene were calculated using dnasp5.

2.6 Phylogenetic and collinearity analyses

Phylogenetic analysis of the mitochondria in this *Camellia* genus was conducted as part of this study. Twenty-five plant mitochondrial genome sequences (17 from the family Theaceae) were downloaded from the National Center for Biotechnology Information database, with the genera *Brassica*, *Aquilaria*, *Dalbergia*, *Hevea*, *Olea*, and *Cocos* as outgroups. Extract the CDS (Coding DNA Sequences) that are shared by 70% or more of the species for the construction of the phylogenetic tree, and multisequence alignment of interspecies sequences was carried out using the mafft software (v7.427, -auto mode). The aligned sequences were joined head to tail and trimmed with trimAl (version 1.4.rev15) (parameter: -gt 0.7) (Capella-Gutiérrez et al., 2009; <https://github.com/inab/trimal>). After trimming, the software jmodeltest-2.1.10 was used for model prediction. The model was determined to be of the general time reversible type. The maximum likelihood evolutionary tree was constructed using the RAxML version 8.2.10 (<https://cme.h-its.org/exelixis/software.html>) software with the GTRGAMMA model selected and bootstrap=1000.

Collinearity analysis was performed using the nucmer (4.0.0beta2) (<https://github.com/mummer4/mummer>) software, with the -maxmatch parameter used for genomic comparison between other sequences and assembled sequences so as to generate dot-plots.

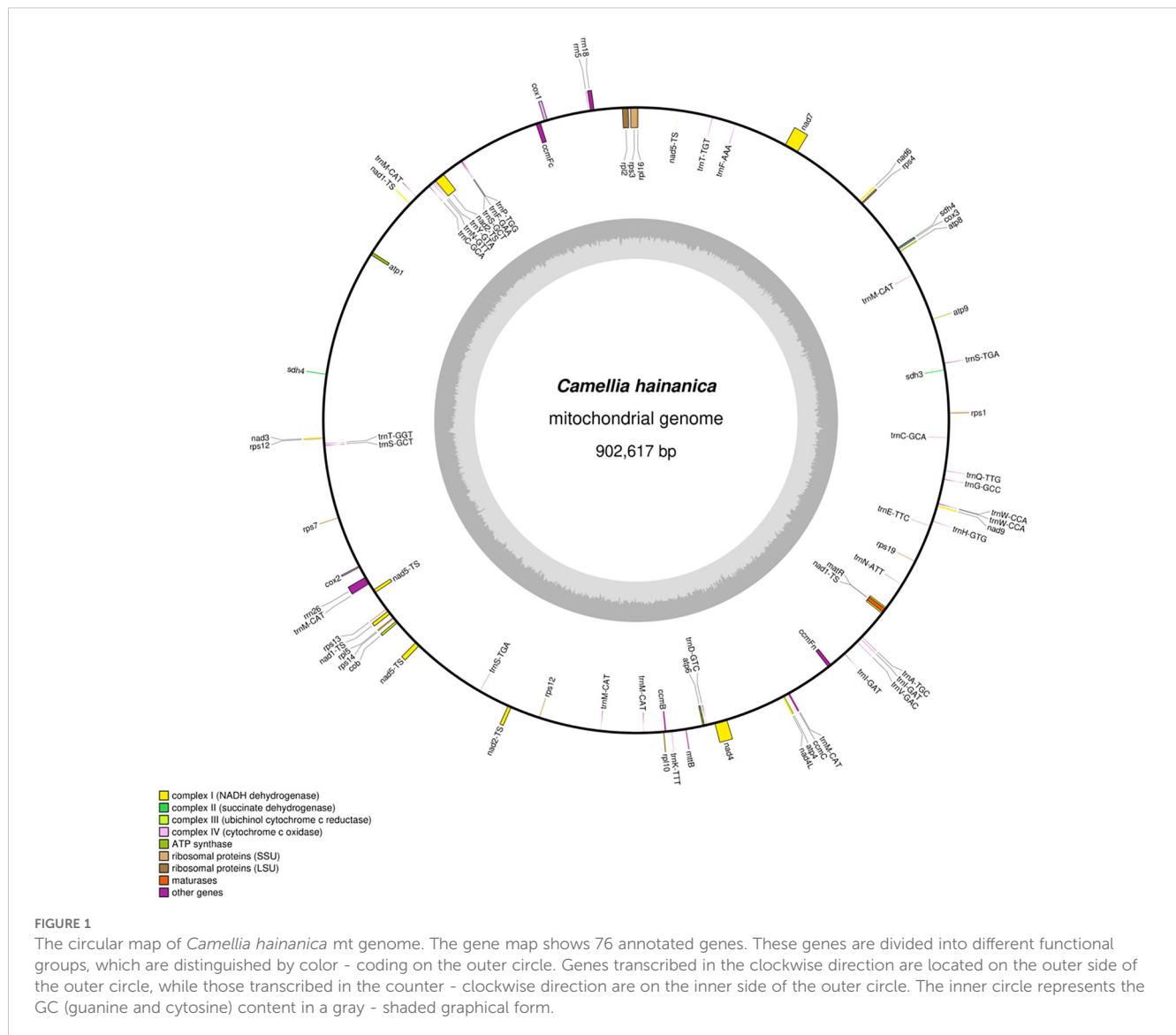
2.7 Analysis of mitochondrial and chloroplast homologous fragments

Homologous sequences between chloroplasts and mitochondria were found using the BLAST software, with the E-value set to 1e-5 and similarity set to not fall <70%.

3 Results

3.1 Decaploid *C. hainanica* mitochondrial genome

In this study, *C. hainanica* was sequenced using the Nanopore sequencing platform, obtaining 9,955,548,123 raw data with a mean sequencing read length of 19,911 bp. The N50 read length was 20,944 bp, and the entire mitochondrial genome of *C. hainanica* had a length of 902,617 bp, with a typical circular structure (Figure 1). The mitochondrial genome of *C. hainanica* comprised 27.18% of A, 27.02% of T, 22.93% of G, 22.86% of C, and 45.79% of GC. The sample was found to contain 76 genes. Specifically, it includes 40 mRNA molecules, with a combined nucleotide sequence length of 32,826 base pairs (bp) and a guanine-cytosine (GC) content of 43.31%. Additionally, there are 32 tRNA molecules, totaling 2,362 bp in nucleotide sequence length and exhibiting a GC content of 49.87%. The sample also comprises three rRNA



molecules, with a combined nucleotide sequence length of 5,650 bp and a GC content of 51.66% as well as one pseudogene.

In the *C. hainanica* mitochondrial genome (Table 1), the *ccmfc*, *rpl2*, *rps3*, *trnA-TGC*, *trnF-AAA*, *trnI-GAT*, *trnS-TGA*, *trnT-GGT*, and *trnT-TGT* genes contain one intron, *nad4* contains two introns, the average length of introns is 744 base pairs (bp). and *nad1*, *nad2*, *nad5*, and *nad7* contain 4 introns, the average length of introns is 1410 base pairs (bp). The genes *rps12*, *sdh4*, *trnC-GCA*, *trnI-GAT*, *trnS-GCT*, and *trnW-CCA* have two gene copies in the *C. hainanica* mitochondrial gene. The gene *trnM-CAT* has six copies in the *C. hainanica* mitochondrial genome.

Plant mitochondrial genes differ significantly in size, gene order, and content, so we selected 17 Theaceae mitochondrial genomes for comparative genomic characterization. Six comparative groups were confirmed, namely *Brassica rapa subsp.*, *B. rapa*, *A. thaliana* (Cruciferae), *Aquilaria sinensi* (Thymelaeaceae), *Dalbergia odorifera* (Leguminosae), *Hevea brasiliensi* (Euphorbiaceae), *Olea europaea subsp.* (Oleaceae), and *Cocos nucifera* (Arecaceae), which were then

studied to obtain the variability of the mitochondrial genomes of decaploid *C. hainanica* (Table 2). The size of the selected mitochondrial genomes ranged from 177,329 to 1,098,121 bp.

3.2 PCG codon usage bias analysis

Codon usage bias can reflect the evolutionary history and environmental adaptations of a species. Three stop codons, UAA, UAG, and UGA, were detected, and C-to-U RNA editing was found in the *ccmFc* gene. The relative synonymous codon usage (RSCU) values of 64 PCGs were also calculated in the *C. hainanica* mitochondrial genome (Figure 2). The 64 PCGs encoded 10,687 codons, including the stop codons. Leu (leucine) was the most common amino acid with 1,097 codons, accounting for 10.2%, followed by Ser (serine) with 996 codons, accounting for 9.3%. The rarest amino acid was Ter (stop codon), with 38 codons, accounting for 0.35%. We found 29 codons with RSCU values >1, of which 27

TABLE 1 Gene profile and organization of the *Camellia hainanica* mt genome.

Group of genes	Gene name
ATP synthase	atp1 atp4 atp6 atp8 atp9
Cytochrome c biogenesis	ccmB ccmC ccmFc* ccmFn
Ubichinol cytochrome c reductase	cob
Cytochrome c oxidase	cox1 cox2 cox3
Maturases	matR
Transport membrane protein	mttB
NADH dehydrogenase	nad1**** nad2**** nad3 nad4** nad4L nad5**** nad6 nad7**** nad9
Ribosomal proteins (LSU)	rpl10 rpl16 rpl2* rpl5
Ribosomal proteins (SSU)	rps1 rps12 rps12 rps13 rps14 #rps19 rps3* rps4 rps7
Succinate dehydrogenase	sdh3 sdh4 sdh4
Ribosomal RNAs	rrn18 rrn26 rrn5
Transfer RNAs	trnA-TGC* trnC-GCA trnC-GCA trnD-GTC trnE-TTC trnF-AAA* trnF-GAA trnG-GCC trnH-GTG trnI-GAT* trnI-GATtrnK-TTT trnM-CAT trnM-CAT trnM-CAT trnM-CAT trnM-CAT trnM-CAT trnN-ATT trnN-GTT trnP-TGG trnQ- TTGtrnS-GCT trnS-GCT trnS-TGA* trnS- TGA trnT-GGT* trnT-TGT* trnV-GAC trnW-CCA trnW-CCA trnY-GTA

*Number of introns

codons (93%) ended in A or U, 1 codon (3.44%) ended in G, and 1 codon (3.44%) ended in C, suggesting that the A/U bias at the third codon is present in *C. hainanica*.

3.3 RNA site editing

In plants, RNA editing is required for gene expression, and C-to-U RNA editing is enriched in mitochondrial and chloroplast genomic

species. In this study, 539 RNA editing sites within 64 PCGs were predicted (Table 3). From analyzing the relationship between gene length and the number of RNA editing sites, it was found that longer coding sequences had more RNA editing sites. However, there is no absolute linear relationship between these factors (Figure 3). From Table 1, it can be seen that amino acid changes occurred at all sites, with the main change patterns being as follows: A (Ala)~V (Val), H (His)~Y (Tyr), L (Leu)~F (Phe), P (Pro)~F, P~L, P~S (Ser), Q (Gln)~*(stop codon), R (Arg)~*, R~C (Cys), R~W (Trp), S~F, S~L, T (Thr)~I (Ile), and T~M (Met). Among these patterns, P~L and P~S had the highest frequency of change, followed by T~M and R~W, while Q~* and R~* had the lowest frequency. The hydrophobicity of 30.43% (164 sites) of amino acids remained unchanged after RNA editing, the hydrophilicity of 12.99% (70 sites) of amino acids remained unchanged after RNA editing, 7.61% (41 sites) of amino acids changed from hydrophobic to hydrophilic, and 48.24% (260 sites) of amino acids changed from hydrophilic to hydrophobic, which therefore led to an increase in the hydrophobicity of the protein. 0.74% (4 sites) of the amino acids changed from hydrophilic to stop codons. Many of the amino acid changes triggered by RNA editing introduce more hydrophobic amino acids into the protein structure, thereby altering the hydrophilicity of the protein and playing a key role in maintaining the regulation of mitochondrial gene expression.

RNA editing can change the PCG start and stop codons. As shown in Table 1, the *cox1* and *nad4L* genes use ACG as the start codon, and so it was hypothesized that this may have been changed by RNA editing. The number of RNA editing sites varied considerably from gene to gene, with the largest number of predictions detected in the cytochrome c biogenesis (*ccmB*, *ccmC*, *ccmFn*, and *ccmFc*), and NADH dehydrogenase (*nad5*) genes.

3.4 SSR analysis

SSRs are characterized by high repeatability, codominant inheritance, uniparental inheritance, and relative conservatism, making them highly efficient molecular markers best suited for species identification and evaluating genetic variation at the population and individual levels.SSRs are stretches of DNA consisting of short unit sequence repeats 1–6 bp in length. Using

TABLE 2 Characterization of the mitochondrial genomes of four species of sect.

	<i>C.hainanica</i>	<i>C. huana</i>	<i>C. lanceoleosa</i>	<i>C.oleifera</i>	<i>C. drupifera</i>
Gene size(bp)	902,617bp	733,752 bp	934,155bp	1,039,838 bp	970,986 bp
Genbank	PV110147	PP975887	PP571817 and PP571818	PP579569	PQ041261 and PQ041262
GC(%)	45.79%	45.73%	45.70%	45.71%	45.73%
Number of gene	76	58	63	63	75
Number of PCGs	40	36	38	38	39
Number of tRNAs	32	20	22	22	32
Number of rRNAs	3	2	3	3	3
Number of pseudo	1				

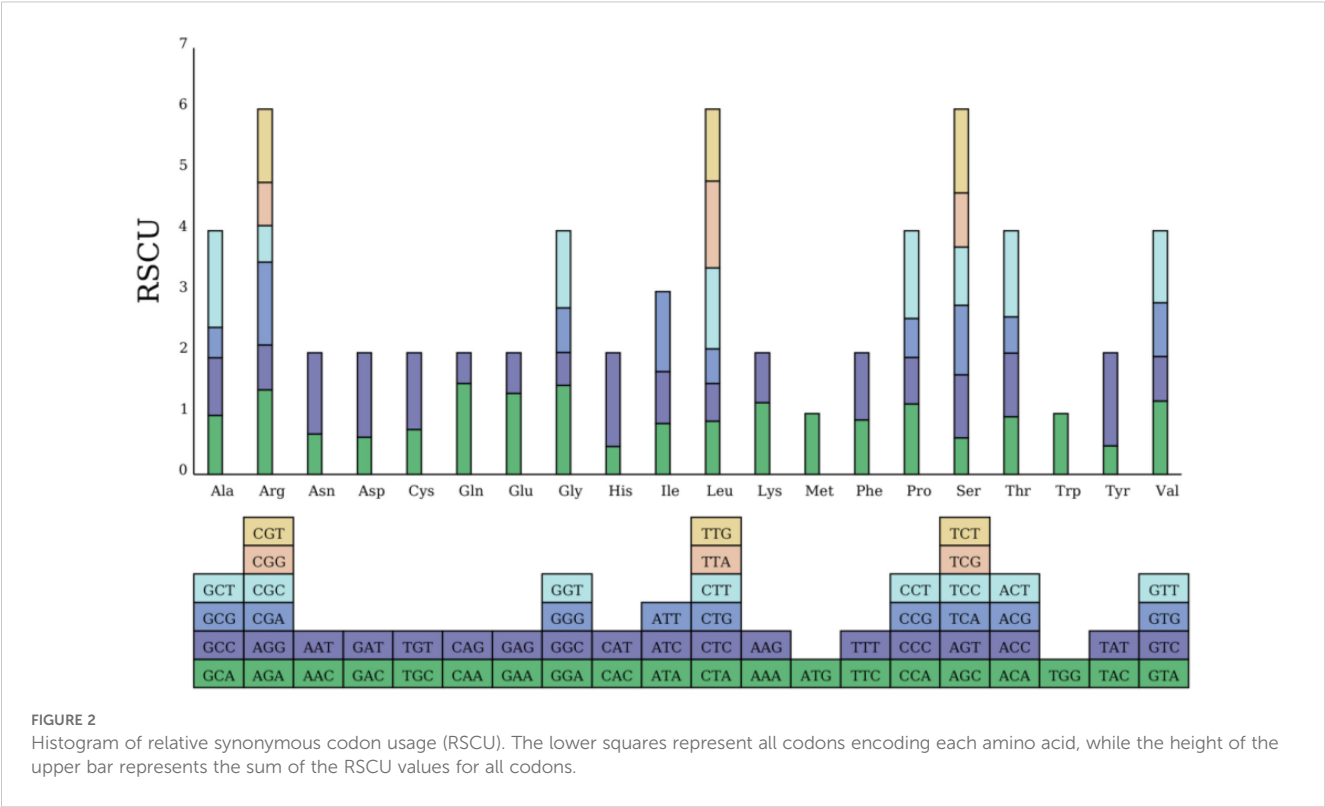


TABLE 3 Statistics regarding the changes in the hydrophilic nature of amino acids induced by RNA editing.

Type	RNA-editing	Number	Percentage
hydrophilic-hydrophilic	CAC (H) => TAC (Y)	9	12.99%
	CAT (H) => TAT (Y)	19	
	CGC (R) => TGC (C)	12	
	CGT (R) => TGT (C)	30	
	total	70	
hydrophilic-hydrophobic	ACA (T) => ATA (I)	4	48.24%
	ACC (T) => ATC (I)	2	
	ACG (T) => ATG (M)	7	
	ACT (T) => ATT (I)	4	
	CGG (R) => TGG (W)	32	
	TCA (S) => TTA (L)	77	
	TCC (S) => TTC (F)	34	
	TCG (S) => TTG (L)	52	
	TCT (S) => TTT (F)	48	
	total	260	
hydrophilic-stop	CAG (Q) => TAG (X)	1	0.74%
	CGA (R) => TGA (X)	3	
	total	4	

(Continued)

TABLE 3 Continued

Type	RNA-editing	Number	Percentage
hydrophobic-hydrophilic	CCA (P) => TCA (S)	8	7.61%
	CCC (P) => TCC (S)	10	
	CCG (P) => TCG (S)	3	
	CCT (P) => TCT (S)	20	
	total	41	
hydrophobic-hydrophobic	CCA (P) => CTA (L)	49	30.43%
	CCC (P) => CTC (L)	8	
	CCC (P) => TTC (F)	8	
	CCG (P) => CTG (L)	36	
	CCT (P) => CTT (L)	22	
	CCT (P) => TTT (F)	14	
	CTC (L) => TTC (F)	5	
	CTT (L) => TTT (F)	12	
	GCC (A) => GTC (V)	1	
	GCG (A) => GTG (V)	7	
	GCT (A) => GTT (V)	2	
	total	164	
	All	539	100%

Type, type of hydrophilicity change; RNA editing, type of RNA editing; Number, number of RNA edits; Percentage, proportion of the sample.

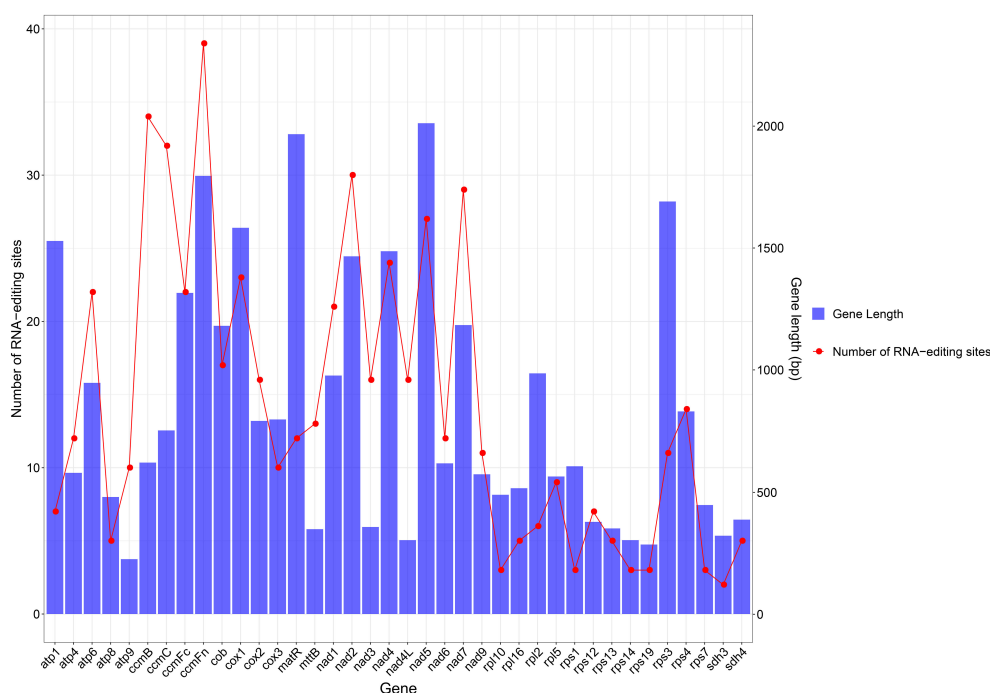


FIGURE 3
Statistics on the number of RNA editing sites per gene.

the MISA tool (Beier et al., 2017), it was found that the minimum number of nucleotide repeats for monomers, dimers, trimers, tetramers, pentamers, and hexamers were 8, 4, 4, 4, 4, and 3 (Zhang et al., 2019), respectively. In this study, a total of 254 SSRs were detected in the mitochondrial genome of *C. hainanica*. The three most common repeat sequences were A/T (93%), AG/TA (44.2%), and AAAG/TTCT (17.3%), with the distribution of these repeat sequences on the genome map shown in Figure 4. Among them, tetranucleotide repeat sequences were the most abundant, with 98 in total, accounting for 38.6% of all SSRs. This was followed by dinucleotides (27.6%), of which there were a total of 70. In addition, there were 29 (11.4%) mononucleotides, 37 (14.6%) trinucleotides (Tri-), 16 (6.3%) pentanucleotides, and 4 (1.6%) hexanucleotides (Table 4). Among the 254 SSRs, dimers and tetramers were the dominant types of SSR motifs, accounting for 66.2% of all detected SSRs. Tandem repeat sequences are the core repeat units of approximately 1–200 bases. As shown in Table 5, 44 tandem repeat sequences with matching of >82% and lengths ranging from 5 to 73 bp were obtained. In addition, there are a total of 691 dispersed repeats exceeding 30 bp, with a total length of 81,690 bp, accounting for 9.05% of the whole mitochondrial genome. The maximum number of repeats ranges from 30 to 65 bp (415 repeats, 60.05%), with three repeats exceeding 1 kb, namely, 16,475, 11,618, and 1,861 bp. The SSR length and number of repeats determine the length and complexity of the repeat base sequences. The above results demonstrate that the *C. hainanica* mitochondrial genome SSR sequences are rich in polymorphisms and can be used for molecular marker development.

3.5 Ka/Ks and Pi analyses

In genetics, the use of Ka/Ks ratios to assess the selective pressure of PCGs during the evolutionary dynamics of similar species is essential for reconstructing phylogenetic relationships and studying the evolution of protein-coding sequences between closely related species. Positive selection ($Ka/Ks > 1$), neutral selection ($Ka/Ks = 1$), and negative selection ($Ka/Ks < 1$) are all possible outcomes. This study analyzed the ratios of Ka and Ks in 40 PCGs present in the mitochondrial genomes of *C. hainanica*, *C. hainanica*, *C. chekiangoleosa* (PP190481.1), *C. lanceoleosa* (PP571818.1), and *C. oleifera* (PP579569.1). The Ka/Ks values of the four common PCGs in *C. hainanica* and *C. hainanica* (NC_086749.1) were zero. The Ka/Ks values of the four common PCGs in *C. hainanica* and *C. chekiangoleosa* (PP190481.1) were also zero. In contrast, only two of the 40 genes common to all species had Ka/Ks values <1, namely *atp6* and *sdh3*, the Ka/Ks values of them are 0.326018 and 0.724737 respectively. This suggests that these two genes have undergone negative selection during evolution, reflecting the tendency of natural selection to remove deleterious non-synonymous mutations.

Nucleotide diversity (Pi) can reveal the magnitude of variation in the nucleic acid sequences of different species, with regions of higher variation able to provide potential molecular markers for population genetics. This study analyzed the Pi of 40 PCGs of the *C. hainanica* mitochondrial genome. Results showed that the Pi values ranged from 0.0002 to 0.00208, with a mean value of 0.000321 (Figure 5). The Pi value of the *sdh3* gene was the largest of these regions at

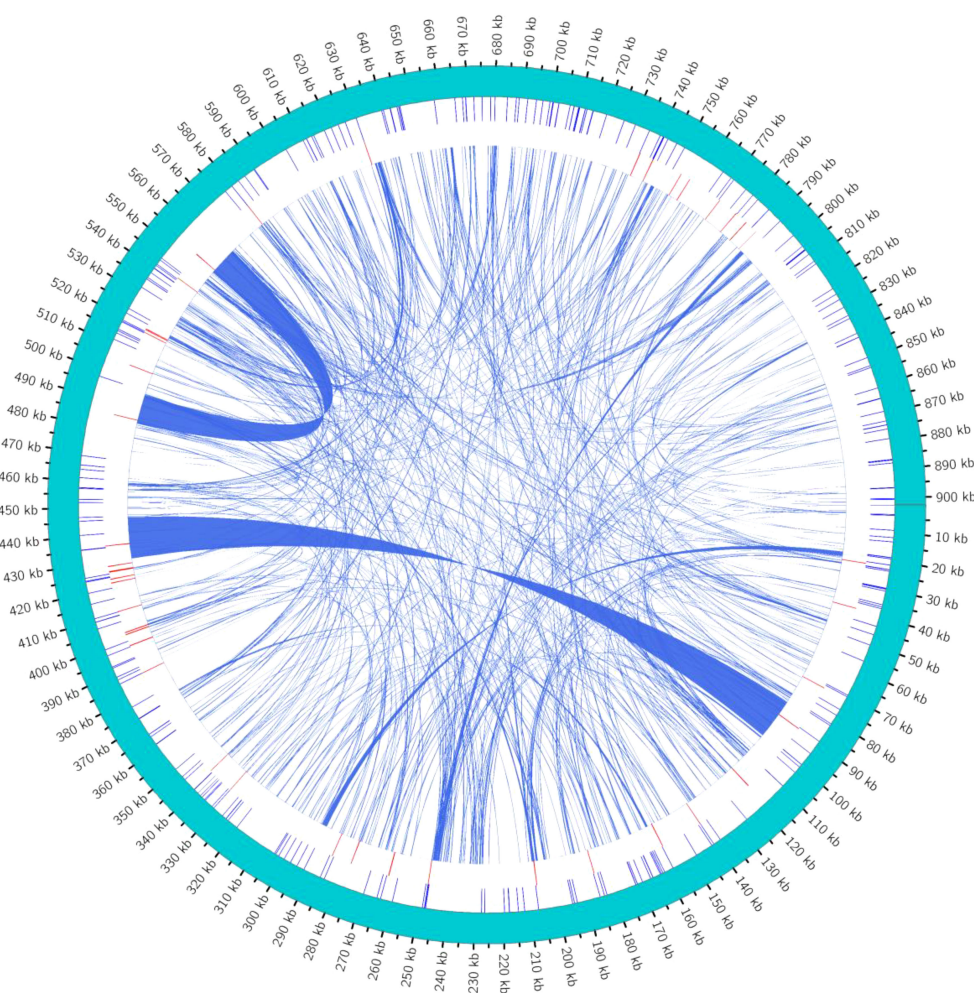


FIGURE 4

Distribution of repeat sequences across the genome. The outermost circle represents the mitochondrial genome sequence. Moving inwards, they are the simple repeat sequences (in blue), tandem repeat sequences (in red), and interspersed repeat sequences in turn. The simple repeat sequences are a type of tandem repeat sequences that are dozens of nucleotides in length and are composed of repeat units consisting of several nucleotides (1 to 6 nucleotides).

0.00374. This suggests that the *sdh3* gene can be used as a molecular marker for the mitochondrial genome analysis of *C. hainanica*, followed by the *atp8* gene with a *Pi* value of 0.00208.

3.6 Phylogenetic and collinearity analyses

This study determined the evolutionary status of the *C. hainanica* mitochondrial genome. A phylogenetic analysis of *C. hainanica* was performed (Figure 6), in which the PCGs were determined to be: *atp1*, *atp4*, *atp6*, *atp8*, *atp9*, *ccmB*, *ccmC*, *ccmFc*, *ccmFn*, *Cob*, *cox1*, *cox2*, *cox3*, *matR*, *mttB*, *nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *rpl10*, *rpl16*, *rpl2*, *rpl5*, *rps1*, *rps12*, *rps13*, *rps14*, *rps19*, *rps3*, *rps4*, *rps7*, *sdh3*, and *sdh4*. The phylogenetic tree was divided into seven groups whose mitochondrial gene sequences were downloaded from GenBank

(<https://www.ncbi.nlm.nih.gov/genbank/>), with the specific genera being *Camellia*(yellow), *Brassica*(green), *Aquilaria*(blue), *Dalbergia* (pale purple), *Hevea*(dark purple), *Olea*(pink), and *Cocos*(orange). Results showed that the species from all families and genera clustered into a single unit, and plants from each different family clustered distinctly with *C. hainanica*. On the phylogenetic tree, the 26 species could be divided into four major branches, with *C. hainanica* and *C. drupifera* clustered closely together, indicating that they had closer phylogenetic affinity.

To better elucidate the conservation of mitochondrial genome evolution between *C. hainanica* and other species in the family Theaceae, a collinearity analysis of the mitochondrial genome sequences was performed (Figure 7). Diploid *C. chekiangoleosa*, diploid *C. lanceoleosa*, tetraploid *C. oleifera*, and octaploid *C. drupifera* were selected for collinearity analysis with decaploid *C. hainanica*. Results showed that the decaploid *C. hainanica* and the

TABLE 4 Frequency of identified simple sequence repeat (SSR) motifs.

SSR-Motif		Reputer/number	ssr number	ssr%
Monomer	A	10/7,11/4,12/2,13/1,16/1(15)	29	11.40%
	C	10/1(1)		
	G	10/1(1)		
	T	10/5,11/4,12/2,15/1(12)		
Dimer	AG	5/13,6/3,7/2(18)	70	27.60%
	AT	5/4,6/3,7/3,8/1,9/1(12)		
	CA	5/1(1)		
	CT	5/9,6/3(12)		
	GA	5/6, 6/2(8)		
	GT	5/1(1)		
	TA	5/9, 6/3, 7/1(13)		
	TC	5/5(5)		
Trimer	AAG	4/1, 5/1(2)	37	14.60%
	AGA	4/2(2)		
	AGC	4/3(3)		
	AGT	5/1(1)		
	CAA	4/1(1)		
	CAT	5/1(1)		
	CTT	4/4(4)		
	GGA	4/1(1)		
	GTA	5/1(1)		
	TAA	4/2(2)		
	TAG	4/2(2)		
	TAT	4/2(2)		
	TCC	4/2(2)		
	TCT	4/5(5)		
	TTA	4/3(3)		
	TTC	4/3(3)		
	TTG	4/2(2)		
Tetramer	AAAT, AAGC, AATA, AATG, AGGA, ATAA, ATTG, CAAA, CAAG, CATT, CCAG, CCGA, CGGG, CTAG, CTAT, CTCC, CTTG, GAAT, GACC, GCTT, GGAG, GGCC, GGCC, GTCC, GTGA, TAGA, TAGC, TATT, TCCC, TCGC, TCTT, TGCT, TGGT, TTAT, TTCA, TTGA	3/1(36)	98	38.60%
	AACA, AAGA, AAGG, ACTA, AGCT, ATAG, ATTC, CAAT, CCTT, CTTC, GAAA, GGAA, TCAT, TGAA, TTCC, TTTG	3/2(32)		
	TTTA	3/3(3)		
	TTCT	3/6(6)		
	AGAA	3/7(7)		
	AAAG	3/11(11)		
	TATC, TCTT	4/1(2)		

(Continued)

TABLE 4 Continued

SSR-Motif		Reputer/number	ssr number	ssr%
	TTCT	5/1(1)		
Pentamer	AATCT,ACTAG,ATAAG,ATAGA,ATATA,CCTAT,CTCTA,CTTTC,GAAAT,GGCTT	3/1(10)	16	6.30%
	CTTTA	3/2(2)		
	TTTAT	4/1(1)		
	ATTAA,ATTAG	6/1(2)		
	TATAA	12/1(1)		
Hexamer	AATAGA,CTATCC,GGATAG,TTGCGC	3/1(4)	4	1.60%

TABLE 5 The tandem repeats analysis of *Camellia hainanica* mitochondrial genome.

NO.	Size	Copy	Repeat sequence	Percent matches	Start	End
1	24	3	CCGGCGCAGGCTCAGCAGGAGGGG	97	21887	21958
2	22	2	AGGTTCGGCTAGCTTAGCTATT	86	39259	39302
3	29	1.9	CTTGCATGGACTGAAAGGCTTCCCCTTTA	88	71394	71449
4	26	2	CTTAGGACATACCCAGGCTAATATGA	92	89408	89459
5	61	2	GGTTTTTCAACGTACGATAGCACGGGTTAGCTTGCTTATTTAGAACTAGTGTCTTAGTTCT	93	117884	118009
6	16	1.9	AGGGTTGTAGAAGTAC	93	140972	141002
7	29	2.2	AACTACCTAGCTACAGGAGGAGAACTACAA	83	157244	157307
8	13	2.3	CTTTCCITTCTATAG	89	184967	184998
9	24	3	CCGGCGCAGGCTCAGCAGGAGGGG	97	206826	206897
10	18	2.1	TCATATTGATTCTATTTT	90	247687	247723
11	18	2.3	TTGAACTGATTCTGAATCC	82	262819	262859
12	20	2	GAAGGGAAGATACCATCCTA	90	277599	277638
13	15	2.1	TCATAGCCGCGAGAGC	88	285307	285339
14	14	2.7	TTCTATTATACTTC	82	331031	331069
15	39	2	AATATCATGATCGGGTCGACCAGGCCAGATCATGAGTGA	97	341180	341258
16	5	12	TATAA	100	386055	386114
17	17	2.1	GAACCCGGCCAGTAAGC	88	397284	397317
18	27	2.4	CCCTTTGAGTAATCTTTAGAATAAAAT	97	401709	401772
19	73	2	GCTTGAACCGTGTACATCGTACGTTGAAAACCCGTGCCATCGTACGTTGGTTCAAGTCTGGTAATGGCGGAAGA	97	402502	402649
20	18	2.1	TTCGATTGGCCTTTCACC	90	410843	410880
21	12	2.1	ATAGGTTCTGAAG	100	421825	421849
22	22	2	CGAAGCCTAGAACCACTGATGA	95	423272	423315
23	12	5.5	TTCTTCGTCCCT	79	425930	425992
24	21	3.7	CTTCTGCTTCTTCGGCCCTTT	80	425923	426003
25	33	2.7	TCATATTCTTTCAATATCTTGTCAATCCTCTCC	84	426116	426205
26	21	2	CGTTTTCTTTTAGAATTGTCT	95	428130	428172

(Continued)

TABLE 5 Continued

NO.	Size	Copy	Repeat sequence	Percent matches	Start	End
27	26	2	CTTAGGACATACCCAGGCTAATATGA	92	435838	435889
28	16	2.4	GATTCCCTTCCGCTAT	91	485472	485509
29	17	2.6	GGAGTACGAGCTTCGAA	96	504996	505039
30	5	6	ATTAA	100	518373	518402
31	26	2	CAATAGAGAAAAGAGGTGTCTGGTGAT	88	519478	519529
32	17	2	AGCGGATCAAAATCGTTG	94	519793	519827
33	15	2.4	AAGTTGCAGAGAGC	90	542376	542408
34	16	2.4	GATTCCCTTCCGCTAT	91	553627	553664
35	6	7.7	TTGCGC	85	580075	580120
36	17	2.2	CTTCTCCTTACTTGGCAG	85	630186	630223
37	20	2	AACCCATTATAATACTAAAG	84	736003	736041
38	5	6	ATTAG	100	741402	741431
39	16	1.9	GTTTGATAGTCTATTCTG	88	753255	753287
40	15	2.5	TAAGAAGAGTAACAG	100	757039	757075
41	19	3.2	ATTCGTTCTCTAGAAGAATG	97	770740	770799
42	14	2.1	TCAGCCCTACAAAG	93	778607	778635
43	33	1.9	GCAACTCCAAATCATGGGGACGAATCCCCCGA	90	783617	783679
44	18	2.1	TCTTTTCTATTAGATTAG	89	788660	788696

octaploid *C. drupifera* had a longer diagonal and good collinearity at the mitochondrial structure level, with a collinearity value of 97.22%, suggesting that the genomes were relatively conserved between the two species in terms of the type, order, and direction of genes, implying that the species share a more recent common ancestor.

3.7 Analysis of mitochondrial and chloroplast homologous fragments

The total length of the chloroplast homologous sequences was 27,468 bp, accounting for 17.5% of the entire chloroplast genome (156,999 bp), while the total size of the mitochondrial homologous sequences was 16,778 bp, accounting for 1.86% of the entire mitochondrial genome (902,617 bp). In total, 34 homologous fragments were found with a total length of 29,841 bp, of which the longest transferred fragment was 9,572 bp, while the shortest fragment was 32 bp (Figure 8). The transfer pathway of the fragments may be first from the chloroplast to the nucleus, and then to the mitochondria. Nine genes were highly similar to mitochondrial genes, namely *trnV-GAC*, *trnI-GAT*, *trnA-TGC*, *trnW-CCA*, *rrn18*, *trnD-GTC*, *trnM-CAT*, *trnN-GTT*, and *ccmC*. These genes may have originated from the mitochondrial genome. Thirty-one genes that were highly similar to chloroplast genes (the 31 genes *rps12*, *rrn4.5*, *rrn23*, *trnA-UGC*, *orf42*, *trnI-GAU*, *rrn16*, *trnV-GAC*, *rpl2*, *rpl23*, *psbJ*, *psbL*, *psbF*, *psbE*, *petL*, *petG*, *trnW-CCA*,

trnP-UGG, *ndhJ*, *ndhK*, *atpE*, *atpB*, *rpoB*, *psbC*, *trnD-GUC*, *trnI-CAU*, *ycf2*, *trnN-GUU*, *trnM-CAU*, *ndhA*, and *psbB*) were likely to have transformed from the chloroplast genome, whereas only partial sequences of these genes were identified in the mitochondrial genome (Table 6). Most of the transferred genes are tRNA genes, which are much more conserved in the mitochondrial genome than in the PCG during evolution.

4 Discussion

C. hainanica is commonly found as a decaploid and octaploid species with large genome data and a complex structure. With advances in sequencing methods, this study obtained more accurate genome assembly sequences of the *C. hainanica* mitochondrial genome (902,617 bp), which has a typical circular structure and is larger than the mitochondrial genomes of most known plants. The GC content was evolutionarily conserved at 45.79%, which is higher than sunflower (45.22%), mango (44.66%), and *Purpuraria* (*Brassica*) (45.23%), all of which are high levels found in higher plants (Makarenko et al., 2021; Niu et al., 2022; Gong et al., 2024). The functional classifications of protein-coding genes within the mitochondrial genome are relatively conserved across species, and their sequences exhibit a high degree of conservation. This suggests that closely related species maintain a high degree of consistency in the composition of their mitochondrial genes. However,

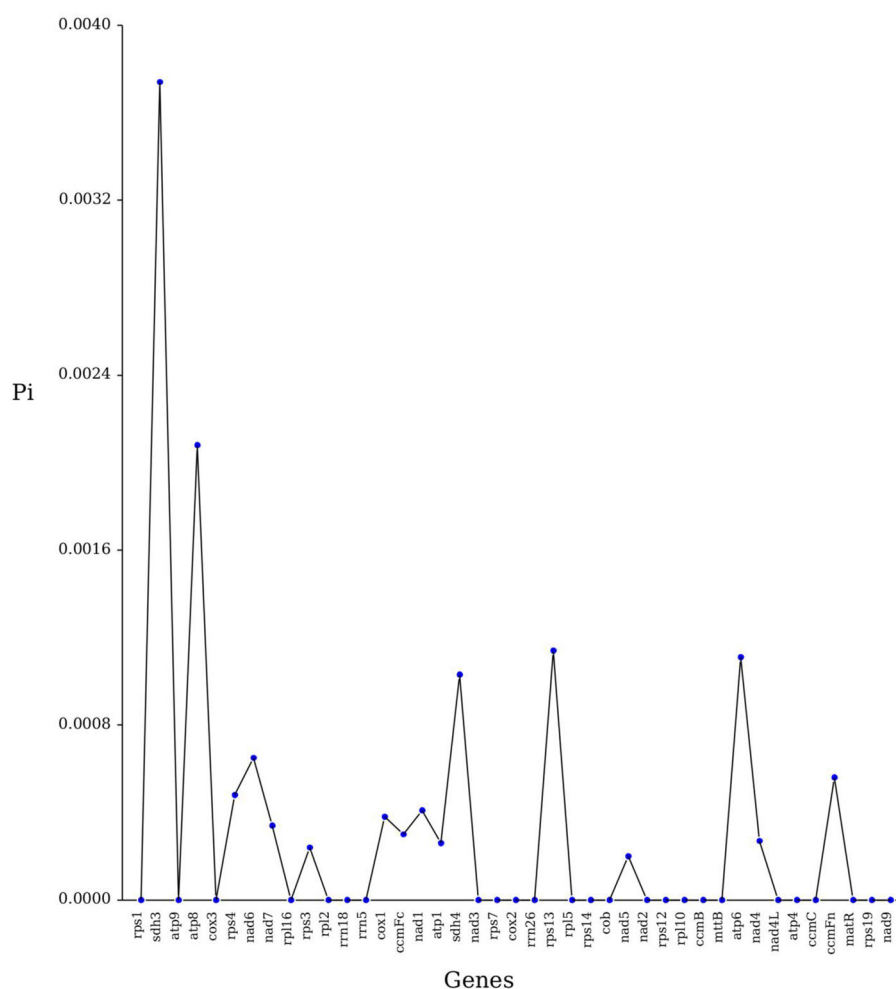


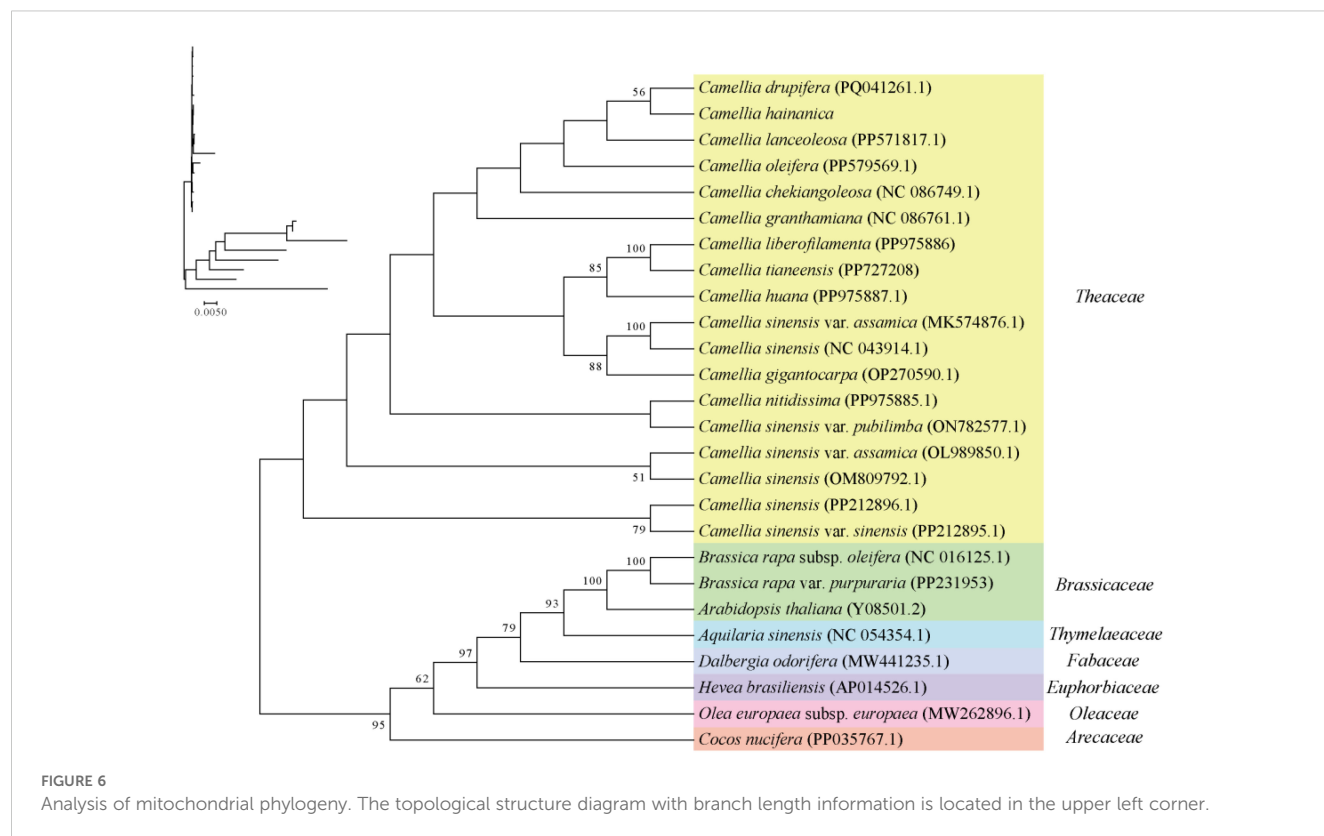
FIGURE 5
Line graph of gene nucleotide diversity (Pi) values.

evolutionary events such as gene rearrangements, losses, or duplications can introduce variability. Consequently, even among species with close genetic relationships, differences in the number of genes and their arrangement within the mitochondrial genome may still be observed (Clifton et al., 2004; Ogihara, 2005). The coding regions of the genome are more conserved than the noncoding regions, and the noncoding regions are also the main source of mitochondrial genome variation (Christensen, 2013). The intergenic region of the mitochondrial genome mainly comprises repeat sequences, chloroplast genome homologous sequences, and contains tandem repeat sequences, dispersed repeat sequences, and SSRs. These are all widespread in the mitochondrial genome (Guo et al., 2017), are essential for the intermolecular recombination of the mitochondrial genome, and are often considered to be the main cause of mitochondrial genome variability (Dong et al., 2018). Most PCGs start with a typical ATN codon (Lin et al., 2017), and some genes contain one or more introns that may play an important role in regulating gene expression.

Ka/Ks ratios are important for assessing the impact of environmental stresses on plants during evolution and can reveal

the effects of genetic changes on the phenotypes of different seed plants. During plant evolution, most mitochondrial genes with Ka/Ks <1 exhibited negative selection, while a few genes with Ka/Ks >1 exhibited positive selection (Xu et al., 2021). It was concluded from the study of the mitochondrial genome of *C. hainanica* that *sdh3* and *atp6* exhibited negative selection, suggesting that these genes may be selected for use in future studies of gene selection and phylogeny in species from the genus *Camellia*. The size and structure of the mitochondrial genome of plants have changed significantly, while functional genes remain conserved. Pi analysis reflects variation in nucleotide sequences between species. Results showed that the Pi value of *sdh3* was the largest among these regions, indicating that the *sdh3* gene can be used as a molecular marker for *C. hainanica* mitochondrial genome analysis.

PCG is usually encoded from the start codon (ATG) to the stop codon (UAA, UAG, and UGA), with the distribution of the amino acid composition found to be consistent with that of *A. thaliana*. Codon usage bias refers to the presence of synonymous codons in a non-random manner across different species (Li et al., 2023a). The analysis of codon usage patterns helps to elucidate the molecular



mechanisms of biological adaptations and to explore evolutionary relationships among species (Ding et al., 2023). Previous studies have shown that there was a bias toward A/U at the ends of codons in plant mitochondrial genomes, with 93% of codons in the *C. hainanica* mitochondrial genome ending in A or U, which may be the result of natural selection, mutational pressure, and genetic drift (Bulmer, 1991). In addition, leucine was found to be the most commonly used amino acid, which is consistent with *Acer truncatum* Bunge (Ma et al., 2022).

The number of RNA editing sites varies from plant to plant and is commonly found in the mitochondrial genomes of gymnosperms and angiosperms. This study obtained 539 RNA editing sites within 64 PCGs of *C. hainanica*, which is lower than that of *Taxus cuspidata* (974) *Ginkgo biloba* (1306) and *Pinus taeda* (1179) (Kan et al., 2020), and much higher than that of *okra* (85) (Li et al., 2022) and *Melia azedarach* L (356) (Hao et al., 2024). The selection of RNA sites in *C. hainanica* showed a high degree of compositional bias. Most RNA editing sites are C-to-U transitions, and most amino acids are converted to hydrophobic amino acids during RNA editing, increasing the hydrophobicity of the edited proteins and thereby increasing the stability of the proteins. Hydrophilic amino acids are distributed on the surface of the protein molecule, whereas hydrophobic amino acids are mainly distributed in the interior of the molecule. The correlation between hydrophilic and hydrophobic amino acids can be used to determine general trends in protein folding. The identification of these RNA editing sites provides important clues for future studies on the evolution of gene function

and the prediction of new codons, and can help to provide a better understanding of gene expression in plant mitochondrial genomes.

Repeat sequences are essential for intermolecular gene recombination and have been widely used to confirm phylogenetic relationships, conduct genetic diversity studies, and achieve species identification due to the high variability and recessive inheritance of SSRs (Ping et al., 2021). The mitochondrial genome of *C. hainanica* contains 254 SSRs, 93% of which are monomers A or T, a genome that is similar to that of sugarcane, *Diospyros kaki* Thunb. 'Taishui' (Ebenaceae), and *Bougainvillea spectabilis* and *Bougainvillea glabra* (Nyctaginaceae) (Yang and Duan, 2024; Zhang et al., 2023). In addition, 44 tandem repeats and 691 dispersed repeats were found in this study, values which are much larger compared to *B. oleracea* var. *Italica* (broccoli) (Zhang et al., 2022a).

Transfer of DNA between the chloroplast and mitochondrial genomes is frequently observed in plant mitochondria (Straub et al., 2013). In higher plants, the size of transferred DNA varies from 50 kb (*A. thaliana*) to 1.1 Mb (*Oryza sativa* subsp. *japonica*; japonica rice) (Smith et al., 2011), depending on the plant species. A total of 29,841 bp of chloroplast DNA was transferred to the *C. hainanica* mitochondrial genome, and around 34 fragments were transferred from the chloroplast genome to the mitochondrial genome, containing nine annotated genes, including seven tRNA genes (*trnV-GAC*, *trnI-GAT*, *trnA-TGC*, *trnW-CCA*, *trnD-GTC*, *trnM-CAT*, and *trnN-GTT*) along with *rrn18* and *ccmC*. The transfer of tRNA genes from the chloroplast to the mitochondrial

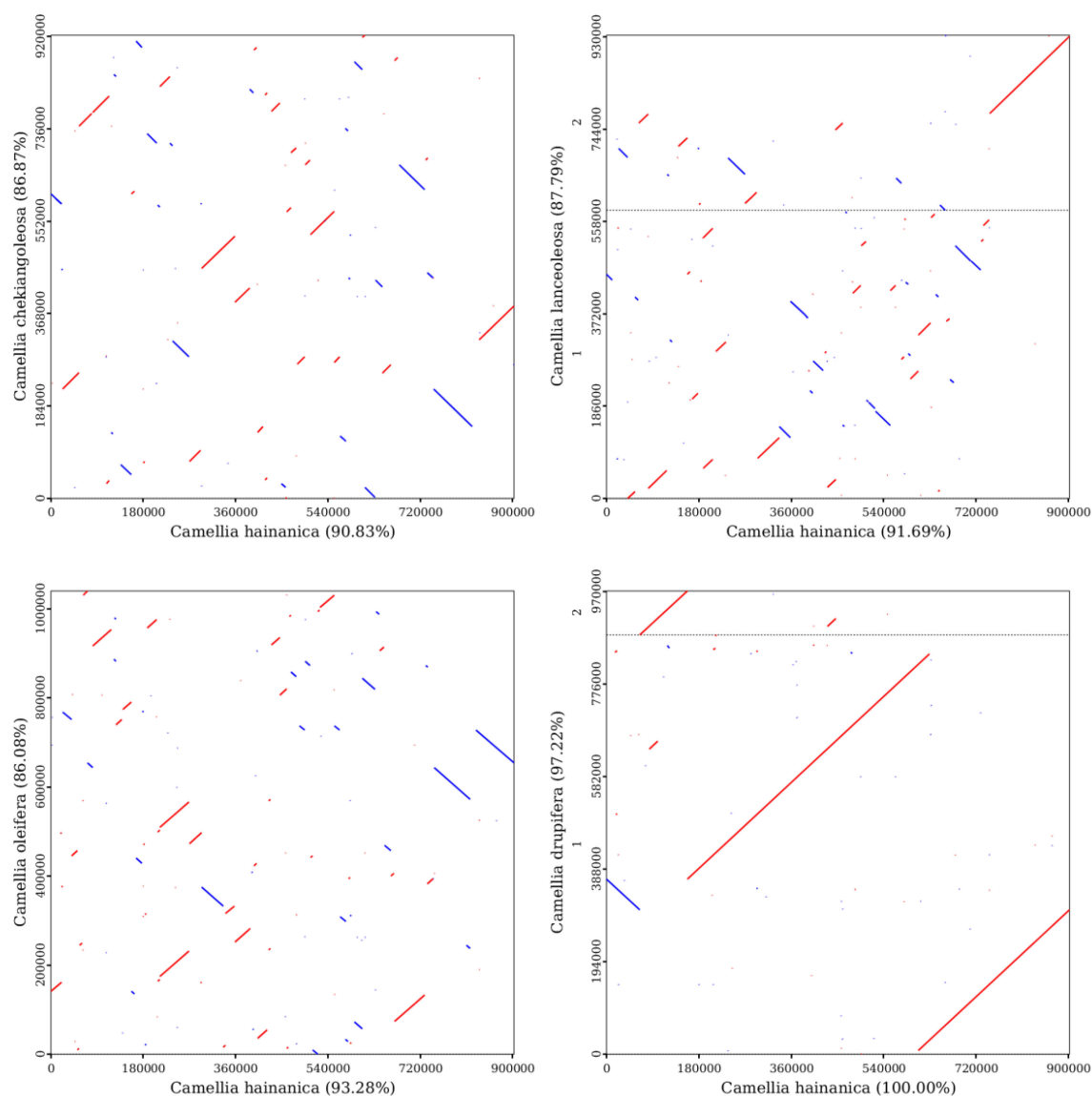


FIGURE 7

Mitochondrial genome collinearity analysis of the four selected *Camellia* species. The red line indicates forward comparison, the blue line indicates reverse complementary comparison, and the black dotted line represents the dividing line between the two chromosomes.

genome is common in angiosperms (Bi et al., 2016). These results are consistent with previous findings, which revealed that tRNA genes are more conserved than PCGs during evolution and that tRNA genes play an integral role in the mitochondrial genome.

In this study, a phylogenetic tree was constructed based on the mitochondrial genomes of 25 plant species, and the whole mitochondrial genome sequence was applied to *C. hainanica* for the first time. The sequenced mitochondrial genome sequences of *Camellia* genus plants and the published mitochondrial genome sequences of six other families were selected for phylogenetic analysis. Results showed that *C. hainanica* was well clustered with the genus *Camellia*, with the classification of the families clearly visible. Of the 25 species, its closest relative was *C. vietnamensis*.

Plant mitochondrial genomes, which are characterized by structural rearrangements, large numbers of genes being lost or gained, mitochondrial to nuclear gene transfer, and very low rates of nucleic

acid mutations, provide unique information for phylogenetic analyses and homologous collinearity analyses among plant mitochondria that can reveal relationships and evolutionary histories among different species. Evolutionary analyses showed that the mitochondrial genome of *C. hainanica* experienced frequent genetic recombination events during the evolutionary process, and these colinear regions not only revealed the conserved patterns of the mitochondrial genome but also reflected the evolutionary relationships and evolutionary history among species, providing a new perspective to reveal the phylogeny and the genetic basis of the *C. hainanica* species.

5 Conclusion

This study sequenced and successfully assembled the complete mitochondrial genome of decaploid *C. hainanica* with a typical

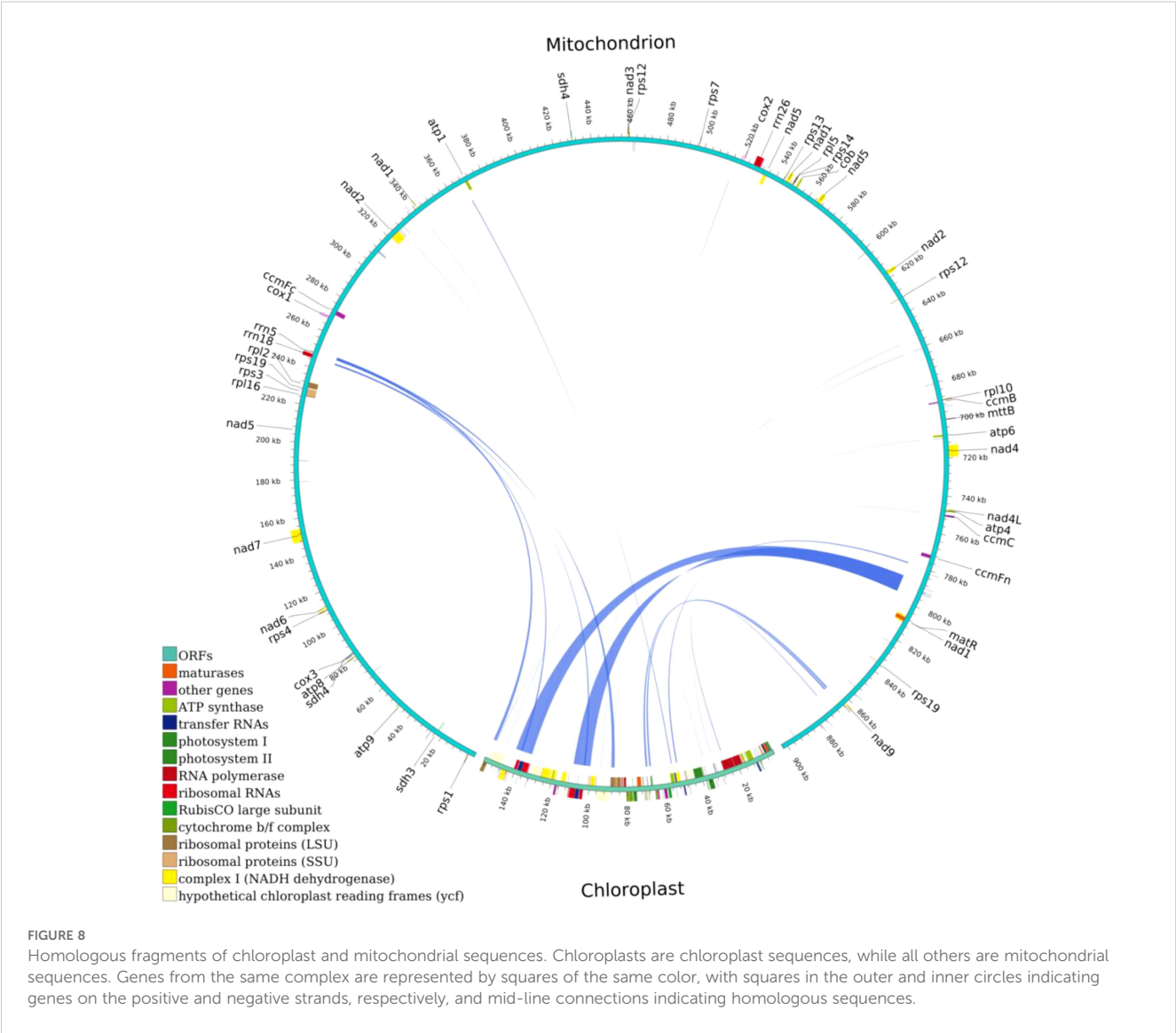


TABLE 6 Chloroplast genome and mitochondrial genome comparison results.

NO	Identical	Length	Mismatches	Gap	Start of alignment in query	End of alignment in query	Start of alignment in subject	End of alignment in subject	Gene(cp)	Gene(mt)
1	99.342	9572	8	7	133762	143315	796675	787141	rps12;rrn4.5;rrn23;trnA-UGC;orf42;trnI-GAU;rrn16;trnV-GAC	trnV-GAC;trnI-GAT;trnA-TGC
2	99.342	9572	8	7	100337	109890	787141	796675	rps12;trnV-GAC;rrn16;trnI-GAU;trnA-UGC;orf42;rrn23;rrn4.5	trnV-GAC;trnI-GAT;trnA-TGC
3	99.467	1689	1	1	155251	156939	249545	247865	rpl23;rpl2	
4	99.467	1689	1	1	86713	88401	247865	249545	rpl2;rpl23	

TABLE 6 Continued

NO	Identical	Length	Mismatches	Gap	Start of alignment in query	End of alignment in query	Start of alignment in subject	End of alignment in subject	Gene(cp)	Gene (mt)
5	87.169	982	82	19	66276	67217	866044	865067	psbJ;psbL; psbF;psbE	
6	85.504	1021	85	32	68281	69295	864681	863718	petL;petG; trnW-CCA; trnP-UGG	trnW-CCA; trnW-CCA
7	81.926	758	99	19	51363	52105	779260	779994	ndhJ;ndhK	
8	86.337	505	38	9	54541	55027	368120	368611	atpE;atpB	
9	81.971	416	41	18	25898	26287	871628	871221	rpoB	
10	74.045	890	174	42	139520	140383	245971	245113	rrn16	rrn18
11	74.045	890	174	42	103269	104132	245113	245971	rrn16	rrn18
12	95.918	147	6	0	36637	36783	340240	340386	psbC	
13	91.275	149	13	0	31883	32031	709660	709808	trnD-GUC	trnD-GTC
14	92.913	127	9	0	154860	154986	660602	660476	trnI-CAU	trnM-CAT
15	92.913	127	9	0	88666	88792	660476	660602	trnI-CAU	trnM-CAT
16	88.889	126	8	2	149246	149371	517689	517570	ycf2	
17	88.889	126	8	2	94281	94406	517570	517689	ycf2	
18	90.517	116	4	1	108480	108595	654611	654503	rrn23	
19	90.517	116	4	1	135057	135172	654503	654611	rrn23	
20	96.429	84	2	1	132349	132431	327451	327368	trnN-GUU	trnN-GTT
21	96.429	84	2	1	111221	111303	327368	327451	trnN-GUU	trnN-GTT
22	97.368	76	2	0	88717	88792	750548	750473	trnI-CAU	ccmC; trnM-CAT
23	97.368	76	2	0	154860	154935	750473	750548	trnI-CAU	ccmC; trnM-CAT
24	93.506	77	5	0	54227	54303	337154	337230	trnM-CAU	trnM-CAT
25	100	42	0	0	122881	122922	787772	787813	ndhA	
26	90.566	53	5	0	150673	150725	590443	590391	ycf2	
27	90.566	53	5	0	92927	92979	590391	590443	ycf2	
28	90.385	52	1	3	155041	155092	80066	80019		
29	90.385	52	1	3	88560	88611	80019	80066		
30	97.297	37	0	1	144192	144228	100452	100417		
31	97.297	37	0	1	99424	99460	100417	100452		
32	97.059	34	1	0	76213	76246	84470	84437	psbB	
33	97.059	34	1	0	76213	76246	430900	430867	psbB	
34	96.875	32	1	0	31989	32020	598206	598175	trnD-GUC	

circular molecular structure. Various genetic aspects of *C. hainanica* were investigated, including its compositional structure, codon preference, RNA editing sites, and repeat sequences, and an integrated alignment analysis was conducted in terms of Ka/Ks and Pi, which revealed the structure of the mitochondrial genome of decaploid *C. hainanica*. Subsequent phylogenetic and collinearity analyses found that decaploid *C. hainanica* was clustered together with *C. vietnamensis* on the phylogenetic tree, suggesting that the two species are more closely related. Horizontal gene transfer of DNA between the mitochondrial and chloroplast genomes was also found in *C. hainanica*, confirming that tRNA genes are genetically conserved over PCGs during evolution. This study provides more comprehensive genetic information on the genome of *C. hainanica*, which is important for revealing the function of the mitochondrial genome and studying the genetic characteristics, evolutionary origin, conservation and utilization, and taxonomic status of this plant family.

Data availability statement

The data presented in the study are deposited in the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov>) repository, accession number PV110147.

Author contributions

SZ: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing-original draft. YZ: Methodology, Software, Formal analysis, Writing - original draft. SL: Methodology, Formal analysis, Writing - review & editing. JG: Software, Resources, Writing - review & editing. HH: Software, Investigation, Data Curation, Writing - review & editing, Funding acquisition. JL: Software, Investigation, Data curation, Writing - review & editing, Funding acquisition. WW: Validation, Investigation, Writing - review & editing. JW: Validation, Investigation, Resources, Writing - review & editing. XH: Validation, Investigation, Resources, Writing - review & editing. HL: Conceptualization, Investigation, Resources, Data curation, Writing - review & editing, Funding acquisition. DH: Conceptualization, Investigation, Resources, Data curation, Writing - review & editing, Funding acquisition.

References

- Andrey, P., Dmitry, A., Dmitry, M., Alla, L., and Anton, K. (2020). Using SPAdes *de novo* assembler. *Curr. Protoc. Bioinf.* 70, e102. Available at: <https://github.com/ablab/spadesmetapv> (Accessed June 10, 2025).
- 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
- Bi, C., Paterson, A. H., Wang, X., Xu, Y., Wu, D., Qu, Y., et al. (2016). Analysis of the complete mitochondrial genome sequence of the diploid cotton *Gossypium raimondii* by comparative genomics approaches. *BioMed. Res. Int.* 2016, 1–18. doi: 10.1155/2016/5040598
- Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129, 897–907. doi: 10.1093/genetics/129.3.897
- Butenko, A., Lukeš, J., Spejger, D., and Wideman, J. G. (2024). Mitochondrial genomes revisited: why do different lineages retain different genes? *BMC Biol.* 22, 15. doi: 10.1186/S12915-024-01824-1
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348
- Christensen, A. C. (2013). Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. *Genome Biol. Evol.* 5, 1079–1086. doi: 10.1093/gbe/evt069
- Clifton, S. W., Minx, P., Fauron, C. M.-R., Gibson, M., Allen, J. O., Sun, H., et al. (2004). Sequence and comparative analysis of the maize NB mitochondrial genome. *Plant Physiol.* 136, 3486–3503. doi: 10.1104/pp.104.044602

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was funded by the National Key Research and Development Program Project (No. 2023YFD2200702), Key Research and Development Program Project of Hainan Province (No. ZDYF2023XDNY055), and Special Project for the Development of High-tech Industries in Hainan Province (No. kjcgh017).

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Correction note

A correction has been made to this article. Details can be found at: [10.3389/fpls.2025.1658310](https://doi.org/10.3389/fpls.2025.1658310).

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.2025.1556379/full#supplementary-material>

- Ding, H., Gao, J., Yang, J., Zhang, S., Han, S., Yi, R., et al. (2023). Genome evolution of *Buchnera aphidicola* (Gammaproteobacteria): Insights into strand compositional asymmetry, codon usage bias, and phylogenetic implications. *Int. J. Biol. Macromolecules* 253, 126738. doi: 10.1016/j.ijbiomac.2023.126738
- Dong, S., Zhao, C., Chen, F., Liu, Y., Zhang, S., Wu, H., et al. (2018). The complete mitochondrial genome of the early flowering plant *Nymphaea colorata* is highly repetitive with low recombination. *BMC Genomics* 19, 614. doi: 10.1186/s12864-018-4991-4
- Gong, Y., Xie, X., Zhou, G., Chen, M., Chen, Z., Li, P., et al. (2024). Assembly and comparative analysis of the complete mitochondrial genome of *Brassica rapa* var. *Purpuraria*. *BMC Genomics* 25, 546. doi: 10.1186/s12864-024-10457-1
- Gu, Y., Yang, L., Zhou, J., Xiao, Z., Lu, M., Zeng, Y., et al. (2024). Mitochondrial genome study of *Camellia oleifera* revealed the tandem conserved gene cluster of nad5–nads in evolution. *Front. Plant Sci.* 15. doi: 10.3389/fpls.2024.1396635
- Gualberto, J. M., Milesheva, D., Wallet, C., Niazi, A. K., Weber-Lotfi, F., and Dietrich, A. (2014). The plant mitochondrial genome: Dynamics and maintenance. *Biochimie* 100, 107–120. doi: 10.1016/j.biochi.2013.09.016
- Guo, W., Zhu, A., Fan, W., and Mower, J. P. (2017). Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. *New Phytol.* 213, 391–403. doi: 10.1111/nph.14135
- Hao, Z., Zhang, Z., Jiang, J., Pan, L., Zhang, J., Cui, X., et al. (2024). Complete mitochondrial genome of *Melia azedarach* L., reveals two conformations generated by the repeat sequence mediated recombination. *BMC Plant Biol.* 24, 645. doi: 10.1186/s12870-024-05319-7
- Kan, S.-L., Shen, T.-T., Gong, P., Ran, J.-H., and Wang, X.-Q. (2020). The complete mitochondrial genome of *Taxus cuspidata* (Taxaceae): eight protein-coding genes have transferred to the nuclear genome. *BMC Evol. Biol.* 20, 10. doi: 10.1186/s12862-020-1582-1
- 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
- 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, C., Zhou, L., Nie, J., Wu, S., Li, W., Liu, Y., et al. (2023a). Codon usage bias and genetic diversity in chloroplast genomes of *Elaeagnus* species (Myrtiflorae: Elaeagnaceae). *Physiol. Mol. Biol. Plants* 29, 239–251. doi: 10.1007/s12298-023-01289-6
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191
- Li, J., Li, J., Ma, Y., Kou, L., Wei, J., and Wang, W. (2022). The complete mitochondrial genome of okra (*Abelmoschus esculentus*): using nanopore long reads to investigate gene transfer from chloroplast genomes and rearrangements of mitochondrial DNA molecules. *BMC Genomics* 23, 481. doi: 10.1186/s12864-022-08706-2
- Li, J., Tang, H., Luo, H., Tang, J., Zhong, N., and Xiao, L. (2023b). Complete mitochondrial genome assembly and comparison of *Camellia sinensis* var. *Assamica* cv. *Dunsta*. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1117002
- Li, Z., Ran, Z., Xiao, X., Yan, C., Xu, J., Tang, M., et al. (2024). Comparative analysis of the whole mitochondrial genomes of four species in sect. *Chrysanth* (*Camellia* L.), endemic taxa in China. *BMC Plant Biol.* 24, 955. doi: 10.1186/s12870-024-05673-6
- Liang, H., Qi, H., Wang, Y., Sun, X., Wang, C., Xia, T., et al. (2024). Comparative chloroplast genome analysis of *Camellia oleifera* and *C. meiocarpa*: phylogenetic relationships, sequence variation and polymorphic markers. *Trop. Plants* 3, 0–0. doi: 10.48130/tp-0024-0022
- Lin, G.-M., Xiang, P., Sampurna, B. P., and Hsiao, C.-D. (2017). Genome skimming yields the complete mitogenome of *Chromodoris annae* (Mollusca: Chromodorididae). *Mitochondrial DNA Part B* 2, 609–610. doi: 10.1080/23802359.2017.1372715
- Lu, C., Gao, L.-Z., and Zhang, Q.-J. (2022). A high-quality genome assembly of the mitochondrial genome of the oil-tea tree *camellia gigantocarpa* (Theaceae). *Diversity* 14, 850. doi: 10.3390/d14100850
- Ma, Q., Wang, Y., Li, S., Wen, J., Zhu, L., Yan, K., et al. (2022). Assembly and comparative analysis of the first complete mitochondrial genome of *Acer truncatum* Bunge: a woody oil-tree species producing nervonic acid. *BMC Plant Biol.* 22, 29. doi: 10.1186/s12870-021-03416-5
- Makarenko, M. S., Omelchenko, D. O., Usatov, A. V., and Gavrilova, V. A. (2021). The insights into mitochondrial genomes of sunflowers. *Plants* 10, 1774. doi: 10.3390/plants10091774
- Niu, Y., Gao, C., and Liu, J. (2022). Complete mitochondrial genomes of three *Mangifera* species, their genomic structure and gene transfer from chloroplast genomes. *BMC Genomics* 23, 147. doi: 10.1186/s12864-022-08383-1
- Ogihara, Y. (2005). Structural dynamics of cereal mitochondrial genomes as revealed by complete nucleotide sequencing of the wheat mitochondrial genome. *Nucleic Acids Res.* 33, 6235–6250. doi: 10.1093/nar/gki925
- Ping, J., Feng, P., Li, J., Zhang, R., Su, Y., and Wang, T. (2021). Molecular evolution and SSRs analysis based on the chloroplast genome of *Callitropsis funebris*. *Ecol. Evol.* 11, 4786–4802. doi: 10.1002/ece3.7381
- Rawal, H. C., Kumar, P. M., Bera, B., Singh, N. K., and Mondal, T. K. (2020). Decoding and analysis of organelle genomes of Indian tea (*Camellia assamica*) for phylogenetic confirmation. *Genomics* 112, 659–668. doi: 10.1016/j.ygeno.2019.04.018
- Sloan, D. B. (2013). One ring to rule them all? Genome sequencing provides new insights into the ‘master circle’ model of plant mitochondrial DNA structure. *New Phytol.* 200, 978–985. doi: 10.1111/nph.12395
- Smith, D. R., Crosby, K., and Lee, R. W. (2011). Correlation between nuclear plastid DNA abundance and plastid number supports the limited transfer window hypothesis. *Genome Biol. Evol.* 3, 365–371. doi: 10.1093/gbe/evr001
- Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., and Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biol. Evol.* 5, 1872–1885. doi: 10.1093/gbe/evt140
- Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: interactive visualization of *de novo* genome assemblies. *Bioinformatics* 31, 3350–3352. doi: 10.1093/bioinformatics/btv383
- Wideman, J. G., Monier, A., Rodríguez-Martínez, R., Leonard, G., Cook, E., Poirier, C., et al. (2020). Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. *Nat. Microbiol.* 5, 154–165. doi: 10.1038/s41564-019-0605-4
- Xie, H., Xing, K., Zhang, J., Zhao, Y., and Rong, J. (2024). Genome survey and identification of key genes associated with freezing tolerance in genomic draft of hexaploid wild *Camellia oleifera*. *J. Hort. Sci. Biotechnol.* 99, 326–335. doi: 10.1080/14620316.2023.2272155
- Xu, Y., Cheng, W., Xiong, C., Jiang, X., Wu, K., and Gong, B. (2021). Genetic Diversity and Association Analysis among Germplasms of *Diospyros kaki* in Zhejiang Province Based on SSR Markers. *Forests* 12, 422. doi: 10.3390/f12040422
- Yang, Y., and Duan, C. (2024). Mitochondrial genome features and systematic evolution of *diospyros kaki* thunb “Taishuu”. *BMC Genomics* 25, 285. doi: 10.1186/s12864-024-10199-0
- Ye, C., He, Z., Peng, J., Wang, R., Wang, X., Fu, M., et al. (2023). Genomic and genetic advances of oiltea-camellia (*Camellia oleifera*). *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1101766
- Zhang, Z. (2022). KaKs_Calculator 3.0: calculating selective pressure on coding and non-coding sequences. *Genomics Proteomics Bioinf.* 20, 536–540. doi: 10.1016/J.GPB.2021.12.002
- Zhang, H., Huang, T., Zhou, Q., Sheng, Q., and Zhu, Z. (2023). Complete Chloroplast Genomes and Phylogenetic Relationships of *Bougainvillea spectabilis* and *Bougainvillea glabra* (Nyctaginaceae). *Int. J. Mol. Sci.* 24, 13044. doi: 10.3390/ijms241713044
- Zhang, F., Li, W., Gao, C. W., Zhang, D., and Gao, L. Z. (2019). Deciphering tea tree chloroplast and mitochondrial genomes of *Camellia sinensis* var. *assamica*. *Sci. Data* 6, 1–10. doi: 10.1038/s41597-019-0201-8
- Zhang, Z., Tao, M., Shan, X., Pan, Y., Sun, C., Song, L., et al. (2022a). Characterization of the complete chloroplast genome of *Brassica oleracea* var. *italica* and phylogenetic relationships in Brassicaceae. *PLoS One* 17, e0263310. doi: 10.1371/journal.pone.0263310



OPEN ACCESS

EDITED BY

Parimalan Rangan,
Indian Council of Agricultural Research
(ICAR), India

REVIEWED BY

Osvaldo Failla,
University of Milan, Italy
Md Shamim,
Bihar Agricultural University, India
Rafael Massahiro Yassue,
GDM, Brazil

*CORRESPONDENCE

Florian Schwander

✉ florian.schwander@julius-kuehn.de

RECEIVED 23 November 2024

ACCEPTED 19 May 2025

PUBLISHED 30 June 2025

CITATION

Heinekamp T, Röckel F, Herzog K, Trapp O,
Töpfer R and Schwander F (2025) Heat
waves reveal additive genetic effects leading
to sunburn resilience of grapevine berries.
Front. Plant Sci. 16:1533345.
doi: 10.3389/fpls.2025.1533345

COPYRIGHT

© 2025 Heinekamp, Röckel, Herzog, Trapp,
Töpfer and Schwander. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Heat waves reveal additive genetic effects leading to sunburn resilience of grapevine berries

Tom Heinekamp, Franco Röckel, Katja Herzog, Oliver Trapp,
Reinhard Töpfer and Florian Schwander*

Julius Kuehn-Institut, Federal Research Centre for Cultivated Plants, Institute for Grapevine Breeding
Geilweilerhof, Siebeldingen, Germany

Grape sunburn is an abiotic stress response induced under heat wave conditions. Heat stress is reaching new dimensions in terms of intensity and frequency in European cool-climate wine-growing regions. The damage to grape berries manifests in browning and shriveling, leading to yield loss and can reduce wine quality. Established management strategies like defoliation of the cluster zone in order to reduce fungal infection pressure could enhance this problem. Climate-adapted cultivars that are resilient to sunburn would resolve those trade-offs in vineyard management. In recent years, grapes grown in the Palatinate wine region of Germany have been affected by sunburn at an unprecedented rate. The intensity of sunburn damage in experimental fields located in this region was assessed for five years, taking advantage of the unexpectedly frequent heat waves in 2019, 2020, and 2022. Phenotyping of the grape sunburn symptoms was carried out in a segregating F1 mapping population of 'Calardis Musqué' x 'Villard Blanc' and a number of varieties. The population consists of 150 genotypes cultivated in two adjacent plots with four plants per F1-individual each, providing sufficient grape material for a reliable evaluation. Composite interval mapping (CIM) using a genetic map and 5 years of phenotypic field data of sunburn damage revealed two strong QTLs located on the lower arm of chromosome 11 with LOD_{max} values of up to 16.3 and 26.1% of explained phenotypic variance and on chromosome 10 with a LOD_{max} value of 10.3 and 14.1% of explained phenotypic variance. The highest sunburn resilience of berries was observed based on an additive effect of a specific allelic combination within both loci. QTL regions were screened for annotated and expressed genes in developing grape berries to provide a first insight into understanding possible principles of sunburn resilience. Some current fungus-resistant varieties (PIWIs), such as 'Calardis Blanc', have demonstrated resilience to sunburn. The reported QTLs open new possibilities to breed for grape sunburn resilient vines using

marker-assisted selection (MAS), but also the challenges are discussed here. This knowledge could facilitate the planting of vineyards with fungus-resistant, sunburn-protected new varieties to ensure yield and wine quality while making viticulture more sustainable.

KEYWORDS

Vitis vinifera L., grapevine breeding, climate change, QTL analysis, abiotic stress, sunburn browning and necrosis, cool climate viticulture, MAS (marker assisted selection)

1 Introduction

Climate change is expected to progress even further worldwide in the coming decades, resulting in unprecedented weather extremes that will challenge the economic viability of agricultural production in every aspect (van Leeuwen et al., 2024). In the last couple of years, weather conditions in Germany have varied considerably between very humid and warm seasons resulting in downy mildew epidemics and hot, dry summers, with alcohol rich vintages, untypical aromas and wine stylistics (Töpfer and Trapp, 2022). The cultivation of traditional grapevine varieties is under increasing pressure due to the loss of fungicidal agents to combat biotic stresses, while at the same time recurrent heat waves and other abiotic stresses result in the loss of yield and typical wine styles in originally cool climate regions. New varieties that are better adapted to abiotic stress induced by the expected change of weather conditions could ensure future production in these areas and contribute to the continued existence of this unique cultural landscape, where wine and tourism are often very important economic factors (Tafel and Szolnoki, 2020).

So far, sunburn resilience of grapes played a minor role in grapevine breeding for cultivars adapted to cool climate. Due to the effects of climate change, in recent years unusual heat waves occurred frequently causing massive sunburn damage in German viticulture (Gambetta et al., 2020). Sunburn is one of the common types of berry shrivel disorders and occurs on fruits exposed to direct sunlight, especially in the heat of the afternoon (Krasnow et al., 2010; Bondada and Keller, 2012; Keller et al., 2016). The berry symptoms start with browning of the epidermis and ranges over necrotic spots to complete berry desiccation. Those are reported to be the results of an combination of excessive photosynthetically active radiation (PAR) and UV radiation as well as high temperature, that can be exacerbated by other stress factors such as water deficit (Gambetta et al., 2020). Gambetta et al., 2020 summarize several damage levels of grape sunburn: (1) degradation of waxes resulting in dehydration, (2) destruction of chlorophyll and cell compartmentalization with subsequent oxidation of polyphenols and browning, (3) cell death in the epidermal layers and exocarp as evidenced by a higher electrical conductivity in berry skin. Grape sunburn damage causes significant yield and quality losses with reduced market value as reviewed in detail

by Gambetta et al., 2020. This findings were recently confirmed by reports of a decrease in must yield of up to 30% and that resulting wines are more yellow colored with oxidative characters (Rustioni et al., 2023; Szmania et al., 2023). Next to the reconsideration of viticultural practices, like canopy management and trellis systems, the development of new grapevine cultivars with improved resilience to sunburn damage is advised, but currently there is a lack of knowledge for an efficient selection in grapevine breeding (Bondada and Keller, 2012; Keller et al., 2016; Delrot et al., 2020). Irrespective of this, significant differences in the resilience against sunburn have already been observed between grapevine cultivars (Müller-Thurgau, 1883; Krasnow et al., 2010; Rustioni et al., 2015). This implies genetically determined parameters identifiable in a classical Quantitative Trait Locus (QTL) analysis based on a F1-crossing population segregating for the trait under examination. Our approach has led to the first QTLs for sunburn resilience in grapevine berries and opens up a perspective for application in marker-assisted selection (MAS). Thus, the work contributes to the increase of breeding efficiency for improved grapevine varieties.

2 Materials and methods

2.1 Plant material and phenotypic evaluation

A biparental F1 mapping population of ‘Calardis Musqué’ (VIVC-No. 4549) x ‘Villard Blanc’ (VIVC-No. 13081) (abbreviated CMxVB) with 150 genotypes was used for the investigation of grape sunburn resilience. The experimental vineyard was established at JKI Geilweilerhof, Siebeldingen, Germany (49°1254’N 8°0248’E) and consists of two plots, Gf-1 and Gf-2, consecutively planted as exact copies with four vines of each genotype within one vineyard grafted on rootstocks of Selection Oppenheim 4 (SO4, VIVC 11473). The vines were planted in 2010 in a vertical shoot positioning (VSP, south north oriented) trellis system and a plant density of 5,000 vines per hectare (2 x 1 m spacing). Pruning is carried out on a flat arch with approx. 10–12 buds per shoot. Parental varieties were co-cultivated in the same plots. Foliage pruning was done 3 weeks after flowering in each season. Defoliation in the cluster zone was carried out later (date indicated

in Figure 1) manually on the eastern side of the canopy, in order to achieve better aeration reducing the risk of mildew and *Botrytis* infections. Sunburn evaluation was conducted in the years 2019 – 2023, when damage became obvious. Additionally, after the heat event in 2019, sunburn assessment data were collected for 83 cultivars from two sets (international and national important cultivars) of the grapevine collection at JKI Geilweilerhof (Table 1). The used weather data was recorded by the station “Siebeldingen (AGM 088)” operated by

the Agrarmeteorologie RLP (<https://www.wetter.rlp.de/>), located 150 m from the experimental plots (Figure 1).

Sunburn damage on berries was phenotyped in the field using a 5-class rating system according to the International Organisation of Vine and Wine (OIV) descriptor 404 for “thermal stress” (1=very low; 3=low; 5=medium; 7=high; 9=very high) (OIV, 2024). For the assessment the rating scheme was modified as followed: 1=no sunburn damage visible; 3=few berries with discoloration and local

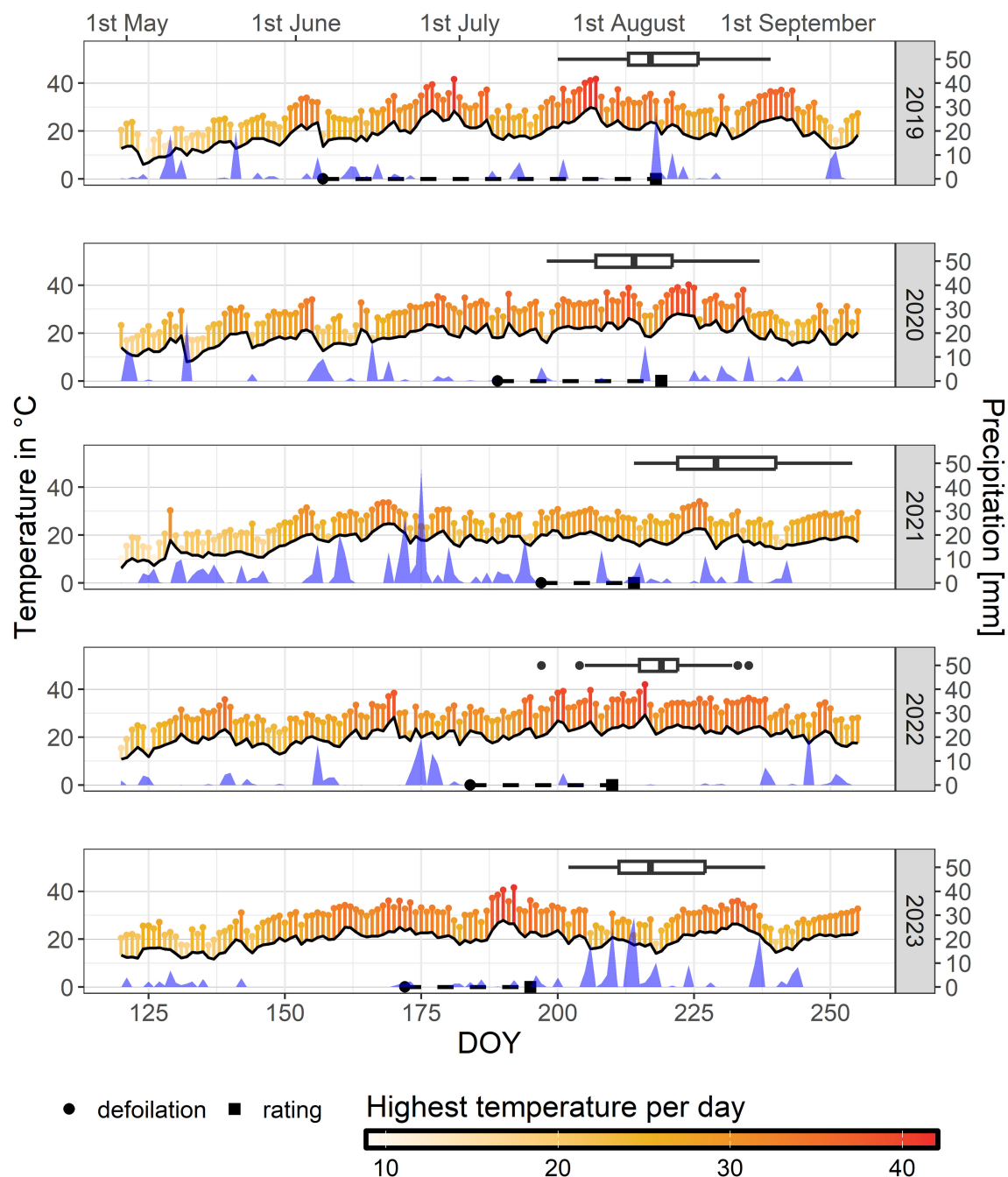


FIGURE 1

Maximum air temperature values (pinheads), mean daytime temperature (black line) 20 cm above ground level and sum of precipitation measured at the weather station “Siebeldingen (AGM 088)” in 150 m distance to the experimental plot for each day of the seasons 2019 to 2023. Dates of defoliation (circle), rating of sunburn damage (square), and dates of veraison (box plot) within the examined mapping population CMxVB are indicated for each year.

TABLE 1 Phenotypic evaluation of sunburn damage in the germplasm repository of JKI Geilweilerhof (DEU098) under field conditions in the year 2019, ranging from 1 = no damage to 9 = massive necrotic damage.

INTERNATIONAL collection	VIVC number	Sunburn	NATIONAL collection	VIVC number	Sunburn	Other plots at JKI	VIVC number	Sunburn
ALIGOTE	312	2	CABERNET MITOS	15499	1	AIREN	157	2
CINSAUT	2672	2	CHARDONNAY BLANC MUSQUE	2456	1	CALANDRO	21797	6
PALOMINO FINO	8888	2	GEWUERZTRAMINER	12609	1	CALARDIS BLANC	22828	2
TEMPRANILLO TINTO	12350	2	SCHEUREBE	10818	1	CALARDIS MUSQUÉ	4549	7
TREBBIANO TOSCANO	12628	2	SILVANER GRUEN	11805	1	FELICIA	20348	2
WELSCHRIESLING	13217	2	DOMINA	3632	2	FELICIA MINIMAL PRUNING	20348	1
ALVARINHO	15689	3	KERNER	6123	2	GF.2004-043-0010	NA	2
ASSYRTIKO	726	3	MUSKATELLER GELB	41578	2	GF.2010-011-0048	NA	3
BARBERA NERA	974	3	OPTIMA	8791	2	PHOENIX	9224	7
CABERNET FRANC	1927	3	RIESLANER	10073	2	REBERGER	19999	3
CHARDONNAY BLANC	2455	3	GEWUERZTRAMINER	12609	2	REBERGER (MINIMAL PRUNING)	19999	2
CLAIRETTE BLANCHE	2695	3	AUXERROIS	792	3	SEYVE VILLARD 39-639	11670	6
GARNACHA TINTA	4461	3	CABERNET DORSA	20002	3	SEYVE VILLARD 39-639	11670	5
TANNAT	12257	3	MUELLER THURGAU WEISS	8141	3	VILLARIS	20347	1
TOURIGA NACIONAL	12594	3	REGENT	4572	3	VILLARIS (MINIMAL PRUNING)	20347	1
VELTLINER GRUEN	12930	3	SOLARIS	20340	3	HEUNISCH WEISS	5374	6
ARAMON NOIR	544	4	PORTUGIESER BLAUER	9620	4			
CABERNET SAUVIGNON	1929	4	BLAUFRAENKISCH	1459	4			
CARMENERE	2109	4	CHARDONNAY BLANC	2455	4			
MERLOT NOIR	7657	4	CHASSELAS BLANC	2473	4			
MUSCAT A PETITS GRAINS BLANCS	8193	4	MORIO MUSKAT	7996	4			
PRIMITIVO	9703	4	RIESLING WEISS	10077	4			
RKATSITELI	10116	4	ACOLON	17123	5			
SAUVIGNON BLANC	10790	4	DAKAPO	14728	5			
SEMILLON	11480	4	ELBLING WEISS	3865	5			
MONASTRELL	7915	5	FABERREBE	4029	5			
RIESLING WEISS	10077	5	PINOT NOIR	9279	5			
ROUSSANNE	10258	5	SAINT LAURENT	10470	5			
SYRAH	11748	5	PINOT BLANC	9272	6			
VIOGNIER	13106	5	PINOT GRIS	9275	6			
CHENIN BLANC	2527	6	PINOT MEUNIER	9278	6			
PLAVAC MALI CRNI	9549	6	PINOT PRECOCE NOIR	9280	6			
CARIGNAN NOIR	2098	7	DORNFELDER	3659	7			
COLOMBARD	2771	7	DUNKELFELDER	3724	7			

(Continued)

TABLE 1 Continued

INTERNATIONAL collection	VIVC number	Sunburn	NATIONAL collection	VIVC number	Sunburn	Other plots at JKI	VIVC number	Sunburn
GAMAY NOIR	4377	7	SCHIAVA GROSSA	10823	7			
PINOT NOIR	9279	7	BACCHUS WEISS	851	8			
DOLCETTO	3626	8						
NEBBIOLO	8417	8						

necrotic sunburn; 5=medium damage with necrotic sunburn defects of a smaller number of berries, 7=a higher number of necrotic damaged berries causing minor yield losses; 9=massive necrotic damage with high yield losses in exposed bunches (Figure 2). One class rating was evaluated for the 4 individuals overall per plot. The data was recorded using the PhenoApp (Röckel et al., 2022).

2.2 Development of locus specific markers

The used SSR marker map is a revised version of the one published by Zyprian et al., 2016. SSR markers were re-analysed to address gaps and uncertainties, and additional markers were included. For fine mapping, primers were deduced based on the grapevine reference genome PN40024 T2T (v5) (Shi et al., 2023) to create locus specific SSR markers on the lower arm of chromosome 11 and the upper arm of chromosome 10. This was done using

WebSat (Martins et al., 2009) and CLC Genomics Workbench (Qiagen, Hilden, Germany).

2.3 Candidate gene identification

The QTL regions, defined by the SSR marker positions flanking the LOD_{max}⁻¹ interval on the PN40024 v5 reference genome, were investigated on sequence level for putative functional genes using Blast2GO provided by Grapedia (grapedia.org, 2024). In addition, positions of these markers in PN40024 v4.2 (Velt et al., 2023) were extracted, to explore the GRape Expression ATlas (Velt, in press) for relevant gene expression studies in relevant organs and within the region of interest. For the v5 version of the reference genome, no studies have been submitted to this database so far. Hence, the v4.1 (Blast2GO) annotation file downloaded from Grapedia was included (Supplementary Table 1). For the screening of the

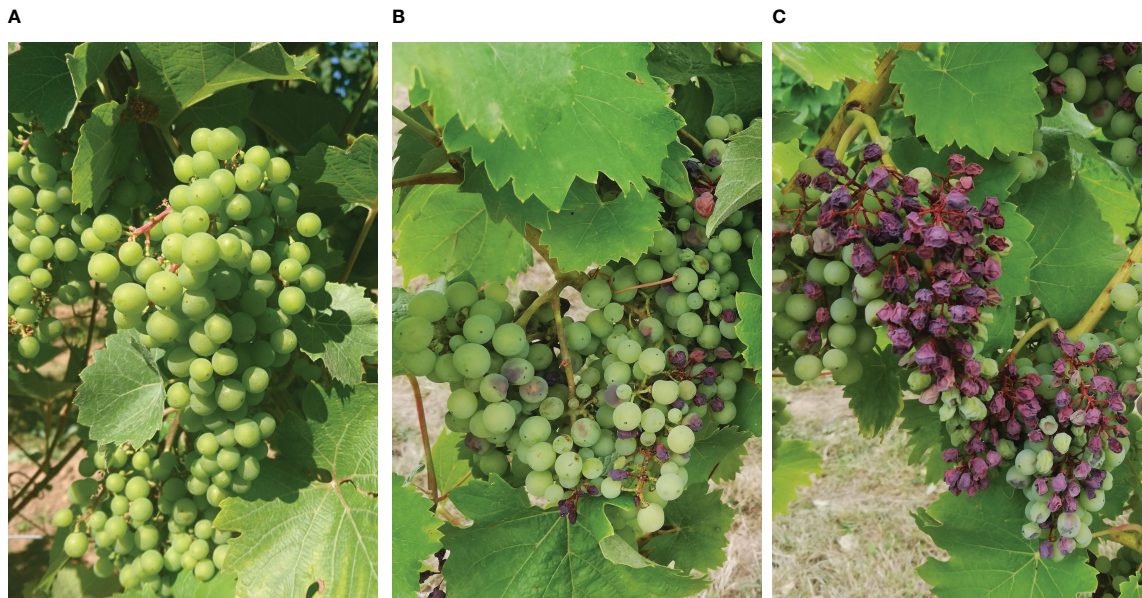


FIGURE 2
Sunburn damage (brown berries) on grapes. The plants show different sensitivity: left – 1=no sunburn damage visible (A); middle – 5=medium damage with necrotic sunburn defects of a smaller number of berries (B), right – 9=massive necrotic damage with high yield losses in exposed bunches (C).

expression studies, only experiments with no treatments and control plants were considered. All cultivars with reports for berry samples within the general developmental stages “Green berries” and “Ripening berries” and extracted organ as “Berries” were included. The data for the extracted QTL regions containing the results of the expression studies are provided in [Supplementary Table 1](#). Subsequently, the genes of interest were compared between v5.1 and v4.1, with consideration given to their activity and putative function. The OneGene platform (<https://onegene-causality-weaver.disi.unitn.it/vitis/network/>) was used to perform network analysis (Pilati et al., 2021) to identify correlating genes.

2.4 Statistical analysis

All analyses were performed via R Statistical Software v4.2.2 (R Core Team, 2022) in RStudio v2024.4.0.735 (Posit team, 2024). Best-linear-unbiased-predictors (BLUPs) (Piepho et al., 2008) for all years were determined with the R-package *phenotype* (Peng, 2020) and the genomic heritability (h^2) was calculated with a simple kinship-matrix from *r/QTL* (Broman et al., 2003) and the package *heritability* (Kruijer et al., 2014). Pearson correlation coefficients with corresponding *p*-values were computed with the R core package *stats*. The genetic map was constructed with *OneMap* (Margarido et al., 2007). Linkage groups were numbered according to the chromosomes of the reference genome. Interval mapping and composite interval mapping in the QTL analysis were performed with *fullsibQTL* (Gazaffi et al., 2014). Cofactors were selected via multiple linear regression function from the *fullsibQTL* package. Interactions between the two QTLs were calculated with *r/QTL*. Additional plots were designed with *ggplot2* (Wickham, 2016).

3 Results

3.1 Examination of heat wave events, viticultural management and date of ratings

Field evaluations were conducted for five consecutive years starting in 2019 to assess a mapping-population for sunburn sensitivity or resilience, respectively. Sunburn events after heat incidents were observable in varying degrees in 2019, 2020, 2022 and were rather weakly expressed in 2021 and 2023. The temperature conditions with mean and maximum temperature per day over the growing seasons within the years of study are shown in [Figure 1](#).

The study was initiated following observations of massive sunburn damaged grapes in early August 2019 and the obvious segregation between the genotypes of the CMxVB population. The conditions in the five years of the study were as follows:

Year 2019: Defoliation of the cluster zone was carried out manually in early June (6th) according to viticultural practices. A natural heat peak at the end of June with a

maximum of 41.6°C on June 30th and a second heat peak exceeding 40°C over three days from 24th to 26th July with a maximum of 41.7°C caused massive sunburn damage at single berries or partly whole bunches. The extent of damage per genotype was rated separately on August 6th for the two repetitive plots at JKI Geilweilerhof ([Figure 3](#)). Differences in the degree of damage between the F1-genotypes were evident and ranged from visually unaffected berries to massive necrotic sunburn symptoms with substantial yield losses.

Year 2020: A deliberately late defoliation on July 7th – 8th in 2020 intended to increase the risk of sunburn, combined with a seven-day heat period at the end of July with temperatures above 32°C and a peak temperature of 38.9°C on July 31, was sufficient to induce sunburn symptoms in sensitive F1-individuals. Segregation within the population was observed ([Figure 3](#)) as documented on August 6th 2020. A second heat event around August 11th occurred after the rating and was not considered in the study.

Year 2021: Defoliation was done on July 16th. No similar high temperatures as in the previous two years were reached in 2021. Maximum temperature occurring during the season was 34.1°C. Therefore, only minor sunburn damage were observed in highly sensitive F1-genotypes. Rating was done at August 2nd (plot Gf-1) and 5th (plot Gf-2).

Year 2022: The first half (Gf-1) of the experimental vineyard was defoliated on July 3rd and 4th at BBCH 73-75, while the second half (Gf-2) was defoliated between July 11th and 14th at BBCH 77-79. An initial heat event occurred in June with temperatures up to 38.4°C, followed by a warm and dry period in the second half of July with temperatures reaching 39.6°C. This episode caused significant damage to sensitive F1-individuals assessed on July 29th. A third heat event around August 4th with temperatures of up to 42°C occurred after the evaluation and is not taken into account.

Year 2023: Defoliation on June 21st and weather conditions with five days above 35°C from July 7th to July 11th and a maximum of 41.6°C on the last day were not sufficient to induce severe sunburn damage. This resulted in a less distinct segregation within the F1 population as evaluated on July 14th and shown in [Figure 3](#).

3.2 Sunburn damage effects

Extend of sunburn damage within the CMxVB population for the individual years of investigation is shown in the histograms given in [Figure 3A](#). By visual examination of the histogram ([Figure 3B](#)) and the QQ-plot ([Supplementary Table 1](#)) based on the BLUP adjusted mean values for the population over all the years, an approximate normal distribution can be assumed, even though this was not confirmed by the Shapiro Wilk test (p-value: 0.001692). The comparison of variation within the F1-population with the

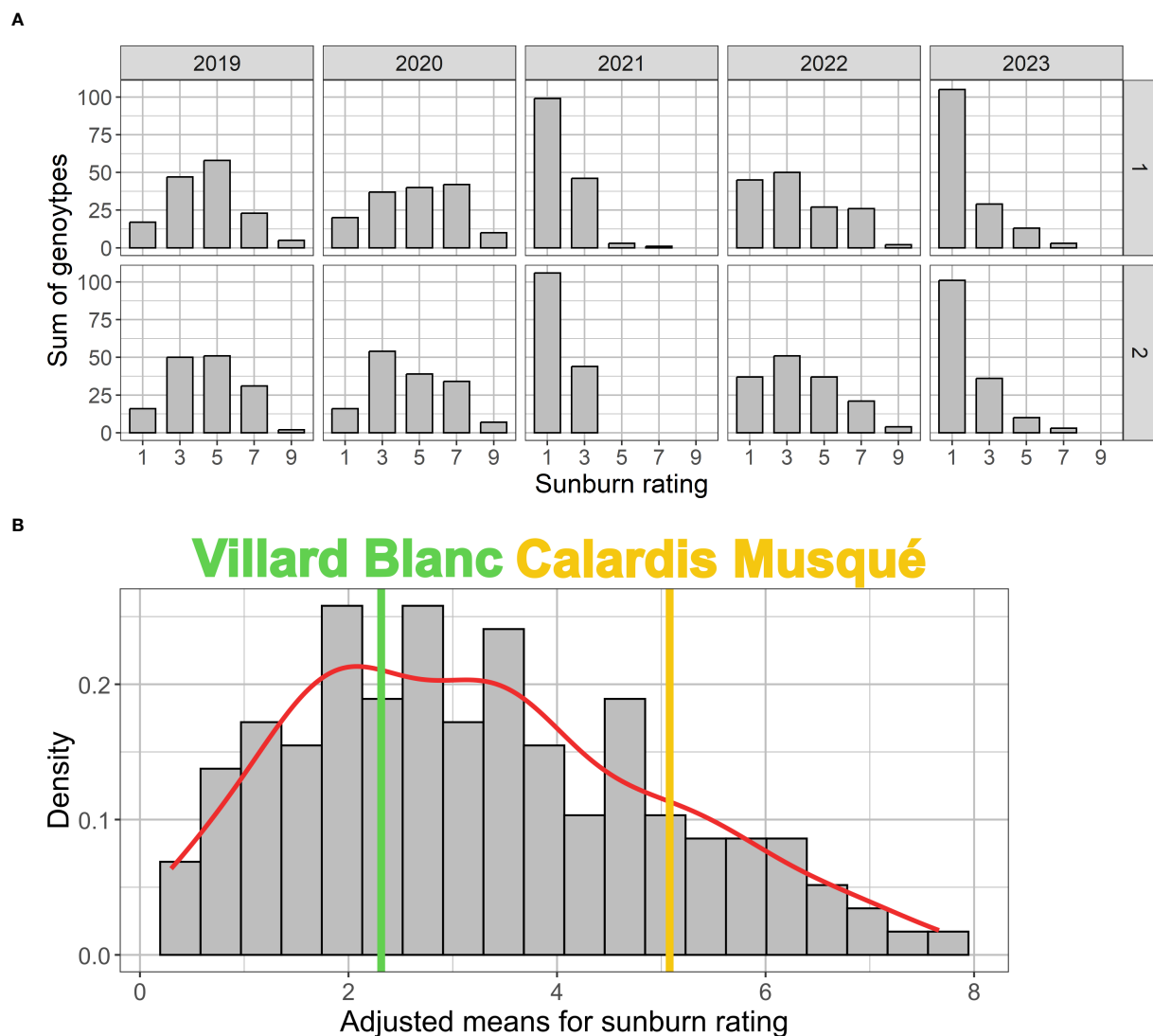


FIGURE 3

Histograms with independent sunburn damage rating in both plots (Gf-1 and Gf-2) of the 'Calardis Musqué' x 'Villard Blanc' F1-population.

(A) Distribution for the five seasons investigated. The rating ranges from "no visible sunburn damage" (=1) to "massive necrotic damage" (=9). The histogram (B) shows the BLUP-adjusted mean for all genotypes over the five years, the corresponding kernel density estimate (red), and the adjusted means for the parental cultivars 'Calardis Musqué' (yellow) and 'Villard Blanc' (green).

observed sunburn resilience of the parents reveals a transgressive segregation, indicating multiple loci to be involved in trait expression.

non-significant p-value = 0.08. This suggests a subordinate influence of the veraison itself on the degree of damage in this population.

3.3 Correlation between sunburn damage and the developmental stage of the grapes

The individuals of the CMxVB population segregate strongly over a period of about 6 weeks for their time of veraison (Frenzke et al., 2024), which is the onset of berry ripening. This means that the grapes of individual genotypes are at different stages of berry development at the time of heat exposure. Therefore, the impact of the developmental stage on the sunburn damage was checked with a Pearson correlation analysis. The day of the veraison combined with the sunburn rating resulted in rather weak correlation of -0.14 and

3.4 QTL analysis of sunburn sensitivity

QTL analyses were performed by using their BLUP value calculated with the rating of four plants over both plots and all five years. Detailed results are presented in [Supplementary Table 2](#). The calculated genomic heritability ranges from 0.29 to 0.65 between the years due to the different characteristics of the heat stress events. The overall BLUP with 0.59 h^2 suggests a strong genetic influence on sunburn resilience.

When performing interval mapping (IM) using the BLUP values, a first QTL on chromosome 11 with a maximum logarithm of the

odds (LOD_{max}) of 9.8 and a second QTL on chromosome 10 with a LOD_{max} of 4.8 were identified (Figure 4). Those QTLs were further confirmed by composite interval mapping (CIM), resulting in a peak on chromosome 11 with a LOD_{max} of 16.3 and one on chromosome 10 with a LOD_{max} of 10.3. The first QTL resulted in 26.1% and the second QTL in 14.1% explained phenotypic variance.

3.5 Construction of the revised genetic map with additional markers

To increase the resolution of the genetic map within the two identified chromosomal regions, 28 SSR-markers were newly developed. Fifteen of those markers segregated in the mapping population and were integrated into the map. A further fine mapping of the lower arm of chromosome 11 based on the given 150 genotypes is restricted due to the lack of recombination events in this region. This resulted in 13 added markers on chromosome 10 and three newly mapped markers on chromosome 11 (Figure 5) compared to the previously published map (Zyprian et al., 2016). Marker positions and primer sequences are attached in Supplementary Table 2. The revised SSR marker map consists of 392 SSR markers on 19 linkage groups with a total length of 1401.5 cM. The average R^2 of the coverage to the grapevine reference genome PN40024 v5sequence is 94.4% for chromosome 10 and 84.5% for chromosome 11.

3.6 Investigation of the genetic interaction between the two loci

Two of the newly developed, integrated, and fully informative SSR markers are located in the respective centres of the two QTLs.

They show the highest LOD values and were therefore selected to study the phenotypic effects. The strength of the mediated sunburn resilience varies depending on the allelic combination within both QTLs (Figure 6). The 16 allele combinations resulting of these two markers, namely GF11-33 (PN40024_v5: chr11, 19,740,009bp) on chromosome 11 and GF10-41 (PN40024_v5: chr10, 5,242,306bp) on chromosome 10 with two distinct alleles each are demonstrated. The interaction plot shows the identified allelic combinations for lowest (Chr10:bd/Chr11:ad) and highest (Chr10:ad/Chr11:bc) sunburn damage as well as their intermediate stages (Figure 6).

In the QTL on chromosome 11 (represented by GF11-33) the allele d, derived from ‘Subereux’ (VIVC-No. 12031) a parent of ‘Villard Blanc’, is associated with sunburn resilience, especially when paired with allele a originating from ‘Bacchus Weiss’ (VIVC-No. 851), a parent of ‘Calardis Musqué’. In the second QTL on chromosome 10 (represented by GF10-41), the allele b, inherited from ‘Seyval Blanc’ (VIVC-No.11558) the second parent of ‘Calardis Musqué’, mediates improved resilience.

3.7 Potential candidate genes for sunburn resilience

The sequence of the reference genome underlying the QTL regions on chromosomes 10 and 11 were investigated for potential candidate genes that may influence sunburn resilience. In PN40024 v5 genome, the flanking markers of the QTL on chromosome 10 (GF10-40 at 4.769 Mb and GF10-46 at 5.650 Mb) cover a region of 880 Mb that includes 108 annotated genes, whereof 97 have reported putative functions. The QTL on chromosome 11 spans a slightly larger region of 1,687 Mb (between the flanking markers GF11-26 at 18.206 Mb and VVMD8 at 19.893 Mb) and includes

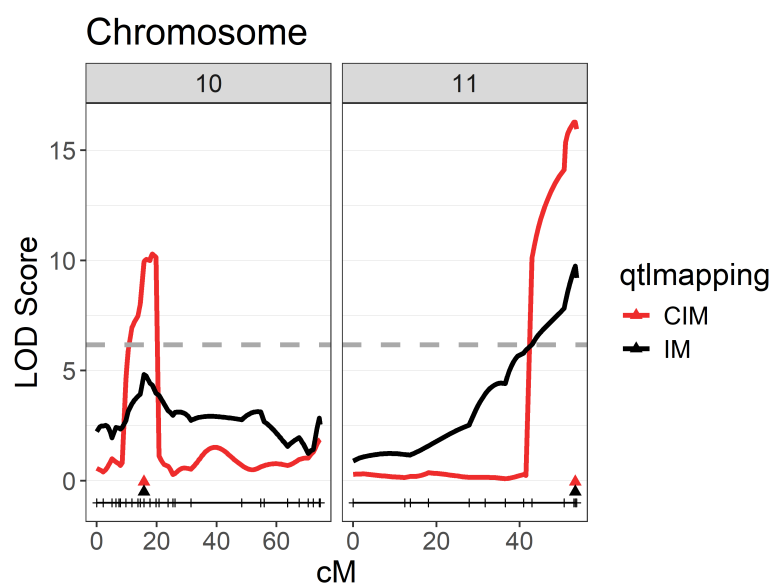


FIGURE 4

QTLs for sunburn resilience on chromosomes 10 and 11 of the mapping population ‘Calardis Musqué’ x ‘Villard Blanc’. QTLs calculated based on the BLUP-values for sunburn damage observed for each F1-genotype in five seasons on two plots from 2019 to 2023. The genome-wide 5% significance threshold for the composite interval mapping with a LOD value of 6.17 is indicated as a dashed line.

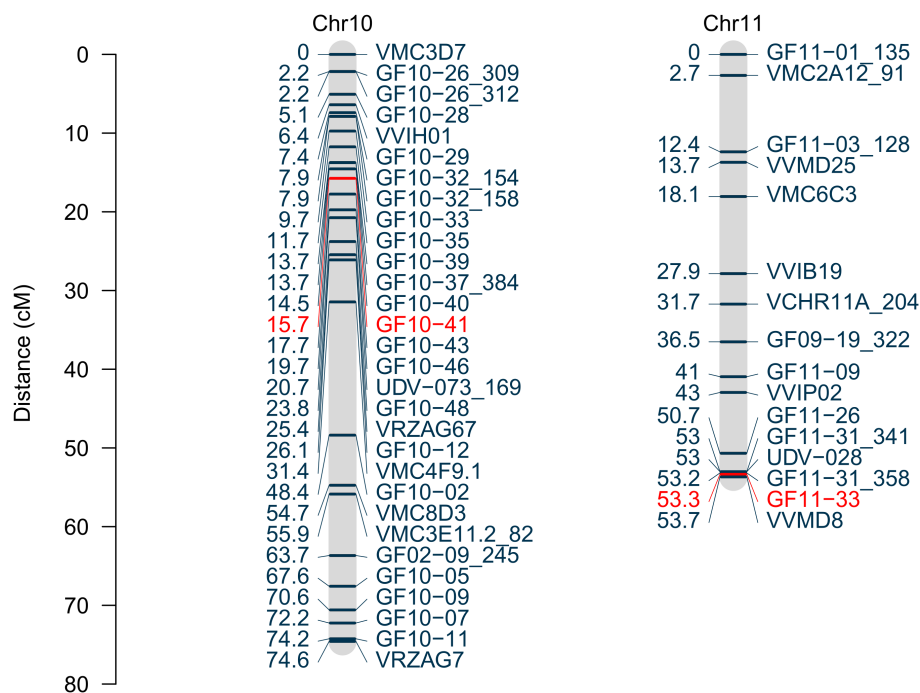


FIGURE 5

Linkage groups representing the chromosomes 10 and 11 based on SSR markers. Genetic positions are displayed in centiMorgan with the according marker names. The LOD peak markers for each QTL are marked in red.

179 annotated genes. 141 of them have reported putative functions (Supplementary Table 1).

Due to missing expression studies referring to the PN40024 v5 genome so far, the QTL regions were additionally transferred via the previously mentioned flanking markers to the reference genome version PN40024 v4.2 (40X). The extracted sequence regions were used to explore the available expression studies (Figure 7). On chromosome 10, the confidence interval $\text{LOD}_{\max}^{-1}$ spans a sequence of 778 Mb (between 4.339 Mb and 5.117 Mb) including 91 annotated genes with reports of 60 putative functions. On chromosome 11, the interval covers the region between 18.102 Mb and 19.775 Mb with a total length of 1.672 Mb, including 138 annotated genes containing 83 with putative functions (Supplementary Table 1). Therefore, an increase of sequence length as well as of annotated genes was observed in the recent reference genome.

To focus on annotations with the highest candidate gene potential, the expression levels in green and ripening berries based on the available data of 13 bioprojects including 12 cultivars were investigated using the GREAT database. The top ten genes in terms of highest measured expression level in one experiment for both developmental stages (green berries and ripening berries) in a cultivar are presented in Figure 7. The widest range in transcripts per million (TPM) between two cultivars at the same stage is 3037 in 'Riesling' compared to 6 TPM in 'Shine Muscat' for *Vitvi11g01268*. While some annotated genes seem to be more or less equally expressed (e.g. *Vitvi10g00372*, *Vitvi10g00397*, and *Vitvi11g01300*) between the cultivars at the same stage, others show a strong variance (e.g. *Vitvi11g01266*, and *Vitvi11g01268*).

3.8 Sunburn damage observed for relevant cultivars

Sunburn symptoms were recorded for two parts of the grapevine collection (national cultivars and international cultivars) at JKI Geilweilerhof and for some selected varieties and breeding lines subsequently to the heat wave of 2019. Under the local weather conditions, as described earlier, those cultivars reacted very differently considering the observed sunburn damage. Very sensitive cultivars like 'Dunkelfelder' (VIVC-No. 3724), 'Dornfelder' (VIVC-No. 3659), 'Schiava Grossa' (VIVC-No. 10823), and 'Bacchus Weiss' (VIVC-No. 851) were found, as well as resilient ones like 'Calardis Blanc' (VIVC-No. 22828), 'Silvaner Gruen' (VIVC-No. 11805), and 'Tempranillo Tinto' (VIVC-No. 12350).

4 Discussion

4.1 Sunburn damage due to heat waves in an historical context

With ongoing climate change, the incidence of grapes damaged by sunburn is increasing. However, grape sunburn events are not an entirely new phenomenon for the German wine-growing regions, as Müller-Thurgau already reported in 1883. At that time, viticulturists had already correctly identified the symptoms of sunburn damage on berries. Müller-Thurgau indicated years, in which moist and cold weather conditions were suddenly followed by hot sunny days as a particular risk for sunburn on unshaded



FIGURE 6

The interaction plot shows the observed phenotypic effects of the allelic combinations on sunburn resilience given by the QTL linked SSR markers on chromosomes 10 (GF10-41) and 11 (GF11-33) in the F1-Individuals of the population CMxVB. Coloration and order depend on the median values of sunburn damaged grapes from low (top/green) to high (below/red). Also shown are the number of F1-individuals (N), phenotypic median, means and standard deviation (Sd) for the individual groups.

grapes. Gambetta et al., 2020 report sunburn incidents in German vineyards in 1892, 1930, 1947, 1966, 1973, 1998, 2007, 2012, and 2019. In addition, the years of 2020 and 2022 continued with sunburn events as reported in the present and further studies (Friedel and Müller, 2022; Waber et al., 2023). Accordingly, in Germany's second largest wine-growing region Palatinate, grapes were affected by unusually high temperature events resulting in remarkable sunburn symptoms in three of the five years between 2019 and 2023, namely 2019, 2020 and 2022. This problem is exacerbated by the common local viticultural practice of extensive defoliation, which many winegrowers perform to improve the microclimate with sun light exposure for aroma formation, rapid drying of grape bunches for health maintenance, improved spray penetration and berry coloration (Drenjančević et al., 2018; Gambetta et al., 2020).

Müller et al., 2023 mentioned, that the likelihood of a berry developing sunburn necrotic symptoms depends on the combination of intensity and duration of the heat event, as well as other biotic and abiotic factors. In the present study sunburn damage in sensitive genotypes was obvious for the years 2019, 2020, and 2022 after hot periods, with peaks exceeding 40°C air temperature (Figure 1).

The symptoms occurred independently of how much time passed since defoliation. In all years, plants could acclimate for at least seven days after leaf removal, which according to Müller et al., 2023 is the period required to adapt the berries to heat events. Thus, the damage occurred despite of adaption and could probably be explained by the much higher, climate change driven recent temperature profiles compared to the very early reports. While, Zschokke, 1930 observed fatal damage to grape berries in 1929 at maximum air temperatures of 32.5°C, while recent air temperature reached above 38.9°C (in 2020) or even 41.7°C (in 2019) resulting in symptomatic berries. The comparison between the historic data and today's bears uncertainties. However, it could be concluded that the higher temperatures in combination with the duration of heat exposure no longer need additional surrounding circumstances such as windlessness, high humidity and low water status to contribute to the same extent to sunburned berries as in former times. Consequently, the sunburn risk has increased substantially.

Another interesting observation by Müller et al., 2023 is that water-stressed grapevines were less or equally sensitive to sunburn necrosis compared with fully irrigated plants. This is in accordance with the observations of Müller-Thurgau (1883) who postulated

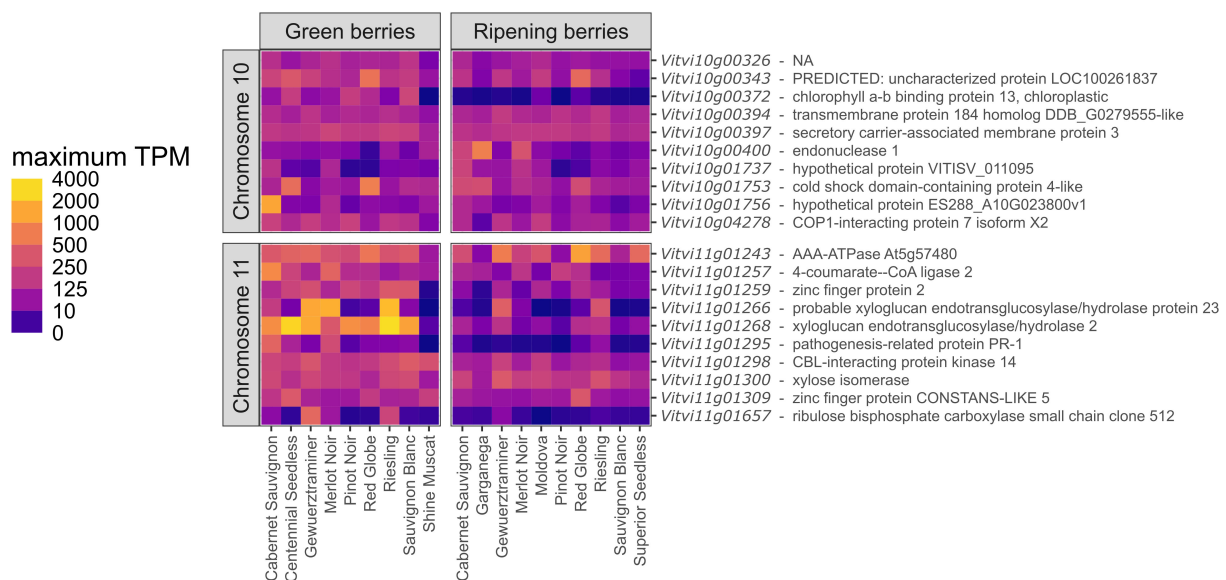


FIGURE 7

The heat map shows a comparison of gene expression levels in transcripts per million (TPM) between different cultivars in the developmental stages "Green berries" and "Ripening berries". The 10 highest expressed genes per QTL region on chromosome 10 and 11 are displayed, summarized through the maximum expression rate per cultivar in one experiment at the same developmental stage. Next to the GeneID, the putative functions with the best hit reported in PN40024 v4.1 are listed.

that water-deficient berries are better protected. On the other hand, he showed cooling effects by berry transpiration (Müller-Thurgau, 1883) and recent studies suggest that the individual water stress level might therefore be of relevance for the expressed sunburn damage symptoms. For example, changes in berry wax composition and transpiration as an result of water stress were reported (Dimopoulos et al., 2020). As all vines in the present study were grafted on the same rootstock variety and planted in the same vineyard, differences between the genotypes would be driven mainly by differential transpiration rates and foliage of the scion genotypes and should be topic of further investigations. Further environmental factors like wind, relative air humidity and global radiation may be additional factors to be considered, but have not shown any statistical significance on sunburn damage here (Supplementary Table 4). In accordance, Gianluca et al., 2025 found out that including radiation data did not improve the accuracy of their machine learning model to predict sunburn damages. A nuanced approach measuring the individual berries, clusters and canopy shading would be necessary, to provide further inside.

Despite the multifactorial stressors that can contribute to the expression of grape sunburn damage, the presented field data over 5 seasons indicate that heat waves have a particularly strong influence on the occurrence of sunburn in sensitive genotypes.

4.2 Influence of the berry ripening on sunburn sensitivity

Gambetta et al., 2020 found reports indicating that the developmental stage affects sunburn sensitivity. While some authors reported an increase in sunburn sensitivity during the berry

development (Webb et al., 2010; Hulands et al., 2013, 2014) others found opposite effects (Müller-Thurgau, 1883; Zschokke, 1931; Müller-Stoll, 1939; Coombe, 1995; Gouot et al., 2019a, b) and it is concluded, that sunburn sensitivity is likely to peak during the veraison, with its major changes in berry-metabolism. The question of developmental dependence of sunburn resilience can be investigated in much more detail and resolution in a suitable cross population that is highly segregating for timing of veraison and berry maturity (Zyprian et al., 2018; Frenzke et al., 2024), than in individual varieties.

Based on the evaluation over five years, including three years of severe sunburn damage, on the given population of 150 genotypes, a clear statement can be made here: Due to a lack of significant correlation between veraison and sunburn damage, the impact of developmental stage is negligible compared to the genotypic effect.

This result is consistent with those of the QTL analysis of the sunburn data, which do not show any QTL for sunburn coinciding with the *Ver1* locus, previously identified as dominant factor for timing of veraison in the same mapping population on chromosome 16 (Frenzke et al., 2024; Zyprian et al., 2016). The results do not contradict the observed changes in sensitivity over the ripening period by other studies, when considering individual grape varieties (Düring and Davtyan, 2002; Greer et al., 2006; Abeyasinghe et al., 2019), but show that the effect is of minor importance compared to the varietal variance.

4.3 QTL results and gene expression analysis

This study revealed two important QTLs for sunburn resilience on chromosomes 10 and 11 based on the phenotypic field data for sunburn damage of five years and the SSR marker based genetic

map. To identify potentially relevant genes for sunburn resilience, all annotated genes within the QTL regions were checked for putative functionality, indicated by gene expression. Knowing that the publicly available RNA expression data sets are not perfectly adequate for a comprehensive analysis with missing phenotypes and divergent cultivars studied, the data was used to determine a general genetic activity within the relevant developmental stages. The occurrence of differing expression rates between the cultivars could be potentially related to physiological differences influencing sunburn characteristics. While berry temperatures above 45–50°C have been observed to induce brown and necrotic spots on berry skins, potentially due to oxidative stress (Griesser et al., 2024), no genes with putative corresponding functions were found to be located within the QTL regions. Heat shock proteins or superoxide dismutases known to be involved in heat stress responses from other species (Kotak et al., 2007) are absent. A MYB24-MYC2 complex has been reported as regulator of heat responses and to induce specialized metabolism pathways in grape berry skin (Zhang et al., 2023, 2024). MYB transcription factors in general were shown to play diverse roles in plant abiotic stress responses (Wang et al., 2021). Therefore, a special focus was set on MYB genes within the QTL regions and two genes with homology to MYB36 (*Vitvi11g01283*) and MYB4 isoform X1 (*Vitvi10g00345*) could be identified. The MYB36 gene is reported as a critical positive regulator of cell differentiation and negative regulator of proliferation in the endodermis of *Arabidopsis* roots (Lieberman et al., 2015). Its expression (together with *APX-1*) under a constitutive promoter resulted in enhanced heat tolerance at grain filling/milking stage in wheat (Firdous et al., 2024), showing its general involvement in plant heat stress reaction. In grapevine, differential expression patterns were observed in roots under differing phosphate concentrations (Gautier et al., 2020), but their potential role in sunburned berries remains open. MYB4 is reported as key regulator in UV tolerance by regulating hydroxycinnamate esters with UV sunscreen functionality in *Arabidopsis* (Wang et al., 2021). For grapevine a function as transcriptional repressor of flavonoid structural genes is known (Pérez-Díaz et al., 2016) and could partly explain the poor color phenotype of sunburn damaged dark berried cultivars (Krasnow et al., 2010). All these findings make the MYB transcription factors a promising target for further studies to investigate different sunburn resilience phenotypes as well as the observed additive effects given by the allelic combinations of both QTLs (Figure 6). To check for a potential connection between the two MYB genes on both chromosomes, the OneGene network was used choosing the shared nodes function. A single gene (*Vitvi04g00472*) coding for a protein kinase domain-containing protein was identified to be positively correlated with both MYB genes and seems to be a part of a complex regulatory network.

On chromosome 11, a cluster of 16 genes with a putative protein function reported as xyloglucan endohydrolase and/or endo-transglycosylase (XTH) was identified and appears to be of particular interest (Figure 7). Xyloglucans play an important role in the cell wall structure and can change their properties like the expandability (Kamerling, 2007). Differences in the underlying activity of those genes in the relevant developmental stages could be a modulating factor. XTHs are highly expressed in green berries when relying on the predicted protein functions for *Vitvi11g01266*

and *Vitvi11g01268*. Their expression reaches very high values up to 2016 and 3036 TPM, respectively, as measured in ‘Riesling Weiss’ (VIVC-No. 10077) (Duchene 2018, PRJEB45016; <https://great.colmar.inrae.fr/>). On the other hand, there are low expression rates in e.g. ‘Sauvignon Blanc’ (VIVC-No. 10790) and ‘Shine Muscat’ (VIVC-No. 22688). The very high expression intensity reported for the two XTH annotations is particularly noteworthy, but could partly be an artefact formed by the overall expression level of the 16-gene cluster, as expression studies can hardly distinguish very similar transcripts. Nevertheless, the observed differences in the expression patterns between cultivars could also be an explanation for varietal differences, but this theory needs to be proven by collecting reliable sunburn damage phenotypes of the cultivars studied. During the ripening stage, the expression level of both reported genes decreases in all varieties.

Interestingly, ‘Shine Muscat’ is a table grape with a bright yellow-green pericarp and was reported to be very sensitive for skin browning and therefore grape production includes bagging of bunches to prevent reduced market value. On initial observations, the browning phenotype (called “*Kasuri-shou*”) appears to be visually similar to sunburn browning symptoms (Katayama-Ikegami et al., 2017). Expression studies identified a specific upregulation of a polyphenol oxidase (*VvPPO2*) and two synthase genes to be associated with berry skin browning (Suehiro et al., 2014). While *VvPPO2* location was uncertain in the study of Suehiro et al., 2014, a Blast analysis of the reported primer sequences on v5 of the reference genome in the present study revealed the position of *VvPPO2* on chromosome 10 at 5.661 Mb, only 11 kb downstream of the lower marker flanking the sunburn QTL. In addition, the primer binding sites of the paralog *VvPPO1* are also located in this region on two positions (5.682 and 5.699 Mb), but this gene doesn’t seem to be linked to browning (Suehiro et al., 2014). As silencing of a *PPO* gene resulted in necrotic lesion on walnut leaves (Araji et al., 2014). This knowledge makes *VvPPO2* an additional potential candidate gene to be involved in grape sunburn response. The position slightly outside the $\text{LOD}_{\max}^{-1}$ confidence interval could be explained by the statistical properties of the QTL analysis. Upstream of the QTL on chromosome 10, a SNP slightly exceeding the significance threshold for sunburned leaf was identified in a GWAS analysis based on 279 grapevine cultivars. The SNP was associated with an 15.4 kDa class V heat-shock-protein (Coupel-Ledru et al., 2024) and the underlying gene could serve as a further possible candidate.

Next to responsive reactions, the sunburn resilience mechanism could be based also on a preemptive measure, which shifts the focus back to the XTHs. Xyloglucan polysaccharides are the main hemicellulose group of the primary cell walls in dicotyledonous plants and can comprise up to 20% of the wall dry matter. Xyloglucans play an important role in interlacing the cellulose microfibrils and have been strongly implicated in the regulation of cell wall extension, particularly in conjunction with the enzyme XTH (Kamerling, 2007). A heat stress comparison between seedlings of the two grapevine varieties ‘Shenfeng’ (VIVC-No. 24745) and ‘Shenhua’ (VIVC-No. 24058) showed a significant higher expression level of XTH genes after a 45 °C heat treatment for 3 and 6 h for the thermo-tolerant variety Shenhua (Zha et al., 2020). Another experiment

investigated four sunlight exposure strategies for bunches in viticulture: (1) basal leaf removal at green berry stage, (2) half-leaf and (3) full leaf removal at veraison as well as (4) leaf moving at veraison. The different leaf management strategies all resulted in an up regulation of four XTH encoding genes in not quite ripe berries (He et al., 2020). Overall, these results show that the XTH genes are linked to abiotic stress response, especially heat and light, making the identified gene cluster an interesting candidate to explain improved sunburn resilience. Other preemptive modifications of the cell wall have already been identified to be important, for example epicuticular waxes. As a coating for grape berries, the wax layer was identified to effectively limit sunburn browning in ten white grape varieties (Domanda et al., 2024). Wax layers were found to differ in ultrastructure between varieties and a QTL for impedance of berries as an indirect measure for the assessment of cuticle thickness and permeability was identified on chromosome 11 (Herzog et al., 2021). Given the considerable distance of 12 Mb between the sunburn and the impedance QTLs the loci do not overlap and annotations within the QTLs showing a direct connection to wax formation were not identified. Indirect interactions are a possible option, particularly regarding the reported differing ‘Shine Muscat’ waxy phenotype and identified wax-related genes on chromosome 11 (Zhang et al., 2021). In unaffected berries of Chardonnay (VIVC-No. 2455) the waxes had an intricately arranged platelet structure orientated perpendicular to the surface. With even the slightest symptoms of sunburn, these waxes have lost the crystalline structure and became relatively amorphous (Greer et al., 2006) indicating a dynamic system.

Completing, we have to note, that 20% of the annotated genes within the QTL regions are of unknown function and could play a role in any of the discussed or undiscussed potential sunburn mechanisms. For chromosome 10, four of them are even in the top 10 of highest expressed genes, which could be a reason for not noting all possible candidate genes.

4.4 Transferability to currently relevant cultivars

Differences in response to heat stress between varieties have long been known. Müller-Thurgau (1883) reported that the grapes of over 50 varieties were more or less affected by the heat event of 1883. Heating experiments with grapes in a metal box showed, which temperature levels are required to induce sunburn symptoms in different varieties. In particular, the early ripening varieties ‘Pinot Noir Précoce’ (PNP; VIVC-No. 9280) and ‘Malingre Précoce’ (MP; VIVC-No. 7249) were not affected in the field by the heat event of 1883, from which Müller-Thurgau deduced a maturity dependence. In contrast, in the present study a substantial sunburn damage in PNP (class 6 Table 1) was observed. This indicates the difficulties of reliable phenotyping of cultivars under the rather uncontrolled, multi-factorial and (formerly) rare natural sunburn events in the field. The commonly early ripening phenotype of PNP, MP, and CM is based on the *Ver1* locus (Frenzke et al., 2024) but does not affect sunburn resilience as reported above.

The rather uncontrolled field conditions result in differential ratings for the same cultivars, even under highly similar conditions

in nearby plots as e.g. reported for ‘Pinot Noir’ (VIVC-No. 9279) that was rated with 5 in the national collection and 7 in the international collection within 100 m planting distance and under similar growing conditions, whereas clonal differences cannot be excluded. Within the F1-population, the rating for both plots correlate with 0.65 across all years, according to Pearson. This highlights the interaction between strong genetic predisposition and multi-factorial environmental influences. More controlled experiments could be helpful, but are still impacted by the preconditions of the grape samples. While berries of ‘Riesling Weiss’ (VIVC-No. 10077), ‘Silvaner Gruen’ (VIVC-No. 11805), ‘Elbling Weiss’ (VIVC-No. 3865), and ‘Pinot Noir’ were already damaged at 42°C in Müller-Thurgau’s metal box, PNP and MP were unaffected even under longer duration at 55°C. Müller-Stoll, 1939 reports Gutedel (‘Chasselas Blanc’; VIVC-No. 2473) to be very sensitive, while ‘Riesling Weiss’ and Traminer (‘Savagnin Blanc’; VIVC-No. 17636) also show damage, but to a less extend. In contrast, ‘Silvaner Gruen’ is described as relatively insensitive and remained unharmed under the respective conditions (Zschokke, 1931; Müller-Stoll, 1939). This can be confirmed by the recent damage ratings of those cultivars within the present study, as shown in Table 1. This broader set of cultivars, even though additional replicates of the observations made in 2019 are missing, can serve as an estimation of their individual sunburn resilience.

4.5 Knowledge transfer to breeding

The genetically based sunburn resilience in the investigated F1 population is mainly characterized by the two loci on chromosomes 10 and 11. As reported in Figure 6, the alleles in both loci have additive effects and the complete allelic pattern of the loci has to be considered to identify the genotypes with the best sunburn resilience. This makes transferability out of the population complex and a better understanding of underlying mechanisms and follow up experiments are necessary to develop suitable and reliable selection tools. Marker assisted selection (MAS) based on few linked SSR markers, as successfully applied for the introgression of resistance loci (Di Gaspero et al., 2012; Welter et al., 2007; Schwander et al., 2012; Töpfer and Trapp, 2022; Zender et al., 2017) or to follow single-locus traits like berry color (Doligez et al., 2002; Röckel et al., 2020) or veraison (Zyprian et al., 2018; Frenzke et al., 2024) is not feasible for sunburn resilience yet.

The allele with the highest impact towards sunburn resilience in this population originates from ‘Subereux’ (chromosome 11), a resistance donor of the early French breeding affords. A number of 54 offspring are reported for this genotype in the VIVC database, where next to ‘Villard Blanc’ with 142 own offspring, most are unreleased breeding lines. As many of recent PIWI cultivars are based on this genepool, a broad set of possible crossing partners should be screened and exploited for further knowledge based breeding. In this QTL, the allele of ‘Bacchus Weiss’ assists the sunburn resilience, but the fact that ‘Bacchus Weiss’ itself is highly sensitive to sunburn (Table 1), underlines the complexity of this multi-factorial trait. Within the second QTL on chromosome 10, the most effective allele originates from ‘Seyval Blanc’; also a

resistant French variety. This allele is accessible in 40 reported offspring and further developed breeding lines. Based on the knowledge gained, breeders can consider sunburn resilience as an additional trait, by choosing suitable crossing partners and by integration of the identified markers into their MAS pipelines.

5 Conclusion

The resilience of grapevine berries to sunburn damage is mostly genetically based. This is demonstrated by the investigated F1 population showing a broad segregation for sunburn damage and a heritability of 0.59 h^2 for sunburn resilience based on 5 years of data acquisition. When sunburn damage inducing heat stress conditions occurred, the further environmental impact as well as the ripening stage were found to be of minor importance. The resulting phenotype can be explained by the additive effects in allele combination within the two identified QTLs on chromosomes 11 and 10. Next steps should include further attempts to identify the underlying genes as well as the validation and establishment of MAS markers based on the gained knowledge to assist breeding. To bridge the gap until new varieties selected based on this knowledge reach the market, it can be recommended to cultivate some recent PIWI varieties that were found to have this desired trait already by coincidence, like ‘Calardis Blanc’, or to limit the damage by appropriate canopy management - as already described by Müller-Thurgau back in 1883.

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

TH: Data curation, Writing – original draft, Writing – review & editing, Formal analysis, Investigation, Methodology, Visualization. FR: Data curation, Writing – review & editing, Validation. KH: Writing – review & editing, Methodology. OT: Writing – review & editing, Resources. RT: Resources, Writing – review & editing, Data curation, Funding acquisition. FS: Funding acquisition, Resources, Writing – review & editing, Conceptualization, Project administration, Supervision, Validation, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The authors are grateful to the German Federal Ministry of Education and Research (BMBF) and the Project Management Jülich (PtJ) for the generous funding of the “SelWineQ” project (FKZ 031B0197, 031B0889, 031B1306) within the program “PLANT BREEDING RESEARCH I, II and III”. The plant material and phenotypic data prior to 2015 were obtained with

funding from the German Federal Ministry of Food and Agriculture (BMEL), for which we would like to express our sincere thanks.

Acknowledgments

The authors thank Anna Werner, Moritz Cappel, Claudia Vogel, Ulrike Braun and Silke Hüther for technical support and helpful advices.

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

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.2025.1533345/full#supplementary-material>

SUPPLEMENTARY TABLE 1

Annotated genes under the LOD_{max}^{-1} QTL-Intervals for sunburn on chromosomes 10 and 11 and corresponding RNA-expression studies.

SUPPLEMENTARY TABLE 2

QTL-analysis results, phenotypic, Sequences and positions of developed SSR markers.

SUPPLEMENTARY TABLE 3

Map-summary, coverage of PN40024 v5 and haplotype map of the parents.

SUPPLEMENTARY TABLE 4

Weather data.

SUPPLEMENTARY FIGURE 1

QQ-plot of the BLUP-adjusted means.

SUPPLEMENTARY FIGURE 2

Genetic map; Chromosomes 1-10.

SUPPLEMENTARY FIGURE 3

Genetic map; Chromosomes 11-19.

References

- Abeyasinghe, S. K., Greer, D. H., and Rogiers, S. Y. (2019). The effect of light intensity and temperature on berry growth and sugar accumulation in *Vitis vinifera* 'Shiraz' under vineyard conditions. *Vitis: J. Grapevine Res.* 58, 7–16. doi: 10.5073/vitis.2019.58.7-16
- Araji, S., Grammer, T. A., Gertzen, R., Anderson, S. D., Mikulic-Petkovsek, M., Veberic, R., et al. (2014). Novel roles for the polyphenol oxidase enzyme in secondary metabolism and the regulation of cell death in walnut. *Plant Physiol.* 164, 1191–1203. doi: 10.1104/pp.113.228593
- Bondada, B. R., and Keller, M. (2012). Not All Shrivels Are Created Equal—Morpho-Anatomical and Compositional Characteristics Differ among Different Shivel Types That Develop during Ripening of Grape (*Vitis vinifera* L.) Berries. *AJPS* 03, 879–898. doi: 10.4236/ajps.2012.37105
- Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112
- Coombe, B. G. (1995). Growth Stages of the Grapevine: Adoption of a system for identifying grapevine growth stages. *Aust. J. Grape Wine Res.* 1, 104–110. doi: 10.1111/j.1755-0238.1995.tb00086.x
- Coupeled-Ledru, A., Westgeest, A. J., Albasha, R., Millan, M., Pallas, B., Doligez, A., et al. (2024). Clusters of grapevine genes for a burning world. *New Phytol.* 242, 10–18. doi: 10.1111/nph.19540
- Delrot, S., Grimpel, J., Carbonell-Bejerano, P., Schwandner, A., Bert, P.-F., Bavaresco, L., et al. (2020). "Genetic and genomic approaches for adaptation of grapevine to climate change," in *Genomic designing of climate-smart fruit crops*. Ed. C. Kole (Springer, Cham Switzerland), 157–270.
- Di Gaspero, G., Copetti, D., Coleman, C., Castellari, S., Eibach, R., Kozma, P., et al. (2012). Selective sweep at the Rpv3 locus during grapevine breeding for downy mildew resistance. In *Theor Appl Genet* 142 (2), pp. 277–286. doi: 10.1007/s00122-011-1703-8
- Dimopoulos, N., Tindjau, R., Wong, D. C. J., Matzat, T., Haslam, T., Song, C., et al. (2020). Drought stress modulates cuticular wax composition of the grape berry. *J. Exp. Bot.* 71, 3126–3141. doi: 10.1093/jxb/eraa046
- Doligez, A., Bouquet, A., Danglot, Y., Lahogue, F., Riaz, S., Meredith, P., et al. (2002). Genetic mapping of grapevine (*Vitis vinifera* L.) applied to the detection of QTLs for seedlessness and berry weight. In *Theor Appl Genet* 105 (5), pp. 780–795. doi: 10.1007/s00122-002-0951-z
- Domanda, C., Paradiso, V. M., Migliaro, D., Pappacogli, G., Failla, O., and Rustioni, L. (2024). Epicuticular waxes: A natural packaging to deal with sunburn browning in white grapes. *Scientia Hort.* 328, 112856. doi: 10.1016/j.scienta.2024.112856
- Drenjančević, M., Rastija, V., Jukić, V., Zmaić, K., Rebekić, A., and Schwander, F. (2018). Effects of early leaf removal on volatile compounds concentrations in Cabernet Sauvignon wines from the Ilok vineyards. *Poljoprivreda* 24, 10–17. doi: 10.18047/poljo.24.1.2
- Düring, H., and Davtyan, A. (2002). Developmental changes of primary processes of photosynthesis in sun- and shade-adapted berries of two grapevine cultivars. *Vitis: J. Grapevine Res.* 41, 63–67. doi: 10.5073/vitis.2002.41.63-67
- Firdous, H., Ali, A., Zafar, M. M., Joyia, F. A., Hamza, M., Razzaq, A., et al. (2024). Nuclear integration of MYB36 and APX-1 genes impart heat tolerance in wheat. *Funct. Integr. Genomics* 24, 185. doi: 10.1007/s10142-024-01456-2
- Frenzke, L., Röckel, F., Wenke, T., Schwander, F., Grützmann, K., Naumann, J., et al. (2024). Genotyping-by-sequencing-based high-resolution mapping reveals a single candidate gene for the grapevine veraison locus Ver1. *Plant Physiol.* 196(1), 244–260. doi: 10.1093/plphys/kiae272
- Friedel, M., and Müller, K. K. (2022). Sonnenbrand verstehen und vorbeugen: Neuer Brennpunkt im europäischen Weinbau. *Der Winzer* 78, 6–10.
- Gambetta, J. M., Holzapfel, B. P., Stoll, M., and Friedel, M. (2020). Sunburn in grapes: A review. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.604691
- Gautier, A. T., Cochetel, N., Merlin, I., Hevin, C., Lauvergeat, V., Vivin, P., et al. (2020). Scion genotypes exert long distance control over rootstock transcriptome responses to low phosphate in grafted grapevine. *BMC Plant Biol.* 20, 367. doi: 10.1186/s12870-020-02578-y
- Gazaffi, R., Amadeu, R. R., Mollinari, M., Rosa, J. R. B. F., Taniguti, C. H., Margarido, G. R. A., et al. (2014). A model for quantitative trait loci mapping, linkage phase, and segregation pattern estimation for a full-sib progeny. *Tree Genetics & Genomes* 10, 791–801. doi: 10.1007/s11295-013-0664-2
- Gianluca, A., Filippetti, I., Pastore, C., Sangiorgio, D., Valentini, G., Bortolotti, G., et al. (2025). Prediction of berry sunburn damage with machine learning: Results on grapevine (*Vitis vinifera* L.). *Biosyst. Eng.* 250, 62–67. doi: 10.1016/j.biosystemseng.2024.12.006
- Gouot, J. C., Smith, J. P., Holzapfel, B. P., and Barril, C. (2019a). Grape berry flavonoid responses to high bunch temperatures post veraison: effect of intensity and duration of exposure. *Molecules* 24, 4341. doi: 10.3390/molecules24234341
- Gouot, J. C., Smith, J. P., Holzapfel, B. P., and Barril, C. (2019b).). Impact of short temperature exposure of *Vitis vinifera* L. cv. Shiraz grapevine bunches on berry development, primary metabolism and tannin accumulation. *Environ. Exp. Bot.* 168, 103866. doi: 10.1016/j.envexpbot.2019.103866
- grapedia.org (2024). Annotation file, October 15. Available online at: <https://grapedia.org/files-download/> (Accessed October 15, 2024).
- Greer, D. H., Rogiers, S. Y., and Steel, C. C. (2006). Susceptibility of Chardonnay grapes to sunburn. *Vitis: J. Grapevine Res.* 45, 147–148. doi: 10.5073/VITIS.2006.45.147-148
- Griesser, M., Savoi, S., Bondada, B., Forneck, A., and Keller, M. (2024). Berry shrivel in grapevine: a review considering multiple approaches. *J. Exp. Bot.* 75, 2196–2213. doi: 10.1093/jxb/erae001
- He, L., Xu, X.-Q., Wang, Y., Chen, W.-K., Sun, R.-Z., Cheng, G., et al. (2020). Modulation of volatile compound metabolome and transcriptome in grape berries exposed to sunlight under dry-hot climate. *BMC Plant Biol.* 20, 59. doi: 10.1186/s12870-020-2268-y
- Herzog, K., Schwander, F., Kassemeyer, H.-H., Bieler, E., Dürrenberger, M., Trapp, O., et al. (2021). Towards sensor-based phenotyping of physical barriers of grapes to improve resilience to botrytis bunch rot. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.808365
- Hulands, S., Greer, D. H., and Harper, J. D. I. (2013). The interactive effects of temperature and light intensity on **Vitis vinifera** cv. 'Semillon' grapevines. I. Berry growth and development. *Eur. J. Hortic. Sci.* 78, 249–257. doi: 10.1079/ejhs.2013/4119075
- Hulands, S., Greer, D. H., and Harper, J. D. I. (2014). The interactive effects of temperature and light intensity on **Vitis vinifera** cv. 'Semillon' grapevines. II. Berry ripening and susceptibility to sunburn at harvest. *Eur. J. Hortic. Sci.* 79, 1–7. doi: 10.1079/ejhs.2014/4222611
- J. P. Kamerling (Ed.) (2007). "Comprehensive glycoscience: from chemistry to systems biology," in *Elsevier Science & Technology Books* (Elsevier Science, San Diego). Imprint.
- Katayama-Ikegami, A., Suehiro, Y., Katayama, T., Jindo, K., Itamura, H., and Esumi, T. (2017). Recombinant expression, purification, and characterization of polyphenol oxidase 2 (VvPPO2) from "Shine Muscat" (*Vitis labruscana* Bailey × *Vitis vinifera* L.). *Biosci. Biotechnol. Biochem.* 81, 2330–2338. doi: 10.1080/09168451.2017.1381017
- Keller, M., Shrestha, P. M., Hall, G. E., Bondada, B. R., and Davenport, J. R. (2016). Arrested sugar accumulation and altered organic acid metabolism in grape berries affected by berry shrivel syndrome. *Am. J. Enol. Vitic.* 67, 398–406. doi: 10.5344/ajev.2016.16048
- Kotak, S., Larkindale, J., Lee, U., Koskull-Döring, P., Vierling, E., and Scharf, K.-D. (2007). Complexity of the heat stress response in plants. *Curr. Opin. Plant Biol.* 10, 310–316. doi: 10.1016/j.pbi.2007.04.011
- Krasnow, M. N., Matthews, M. A., Smith, R. J., Benz, J., Weber, E., and Shackel, K. A. (2010). Distinctive symptoms differentiate four common types of berry shrivel disorder in grape. *Cal Ag* 64, 155–159. doi: 10.3733/ca.v064n03p155
- Kruijer, W., Boer, M. P., Malosetti, M., Flood, P. J., Engel, B., Kooke, R., et al. (2014). Marker-based estimation of heritability in immortal populations. *Genetics* 199, 379–398. doi: 10.1534/genetics.114.167916
- Liberman, L. M., Sparks, E. E., Moreno-Risueno, M. A., Petricka, J. J., and Benfey, P. N. (2015). MYB36 regulates the transition from proliferation to differentiation in the Arabidopsis root. *P. Natl. Acad. Sci. U.S.A. P Natl. Acad. Sci. U.S.A.* 112, 12099–12104. doi: 10.1073/pnas.1515576112
- Margarido, G. R. A., Souza, A. P., and Garcia, A. A. F. (2007). OneMap: software for genetic mapping in outcrossing species. *Hereditas* 144, 78–79. doi: 10.1111/j.2007.0018-0661.02000.x
- Martins, W. S., Lucas, D., Neves, K. F., and Bertoli, D. J. (2009). WebSat - A web software for microsatellite marker development. *Bioinformatics* 3, 282–283. doi: 10.6026/97320630003282
- Müller, K., Keller, M., Stoll, M., and Friedel, M. (2023). Wind speed, sun exposure and water status alter sunburn susceptibility of grape berries. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1145274
- Müller-Stoll, W. R. (1939). Studien über hitzebeschädigungen an weintrauben. *Z. für Pflanzenkrankheiten (Pflanzenpathologie) und Pflanzenschutz* 49, 577–589. Available at: <https://www.jstor.org/stable/43229728>.
- Müller-Thurgau, H. (1883). Über Beschädigung von Trauben durch Sonnenbrand. *Der Weinbau* 9, 143–145.
- OIV (2024). 3rd edition of "oiv descriptor list of grape vine varieties and vitis species". (Dijon, France: International Organisation of Vine and Wine).
- Peng, Z. (2020). *Phenotype: A tool for phenotypic data processing*. Available at: <https://CRAN.R-project.org/package=Phenotype>.
- Pérez-Díaz, J. R., Pérez-Díaz, J., Madrid-Espinoza, J., González-Villanueva, E., Moreno, Y., and Ruiz-Lara, S. (2016). New member of the R2R3-MYB transcription factors family in grapevine suppresses the anthocyanin accumulation in the flowers of transgenic tobacco. *Plant Mol. Biol.* 90, 63–76. doi: 10.1007/s11103-015-0394-y
- Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8

- Pilati, S., Malacarne, G., Navarro-Payá, D., Tomè, G., Riscica, L., Cavecchia, V., et al. (2021). Vitis oneGenE: A causality-based approach to generate gene networks in vitis vinifera sheds light on the laccase and dirigent gene families. *Biomolecules* 11, 1744. doi: 10.3390/biom11121744
- Posit team (2024). *RStudio: Integrated Development Environment for R* (Boston, MA: Posit Software, PBC).
- R Core Team (2022). *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing).
- Röckel, F., Moock, C., Braun, U., Schwander, F., Cousins, P., Maul, E., et al. (2020). Color Intensity of the Red-Fleshed Berry Phenotype of Vitis vinifera Teinturier Grapes Varies Due to a 408 bp Duplication in the Promoter of VvmybA1. In *Genes* 11 (8), pp. 891. doi: 10.3390/genes11080891
- Röckel, F., Schreiber, T., Schüler, D., Braun, U., Krukenberg, I., Schwander, F., et al. (2022). PhenoApp: A mobile tool for plant phenotyping to record field and greenhouse observations. *F1000Res* 11, 12. doi: 10.12688/f1000research.74239.1
- Rustioni, L., Altomare, A., Shanshiashvili, G., Greco, F., Buccolieri, R., Blanco, I., et al. (2023). Microclimate of grape bunch and sunburn of white grape berries: effect on wine quality. *Foods* 12, 621. doi: 10.3390/foods12030621
- Rustioni, L., Milani, C., Parisi, S., and Failla, O. (2015). Chlorophyll role in berry sunburn symptoms studied in different grape (Vitis vinifera L.) cultivars. *Scientia Hort.* 185, 145–150. doi: 10.1016/j.scienta.2015.01.029
- Schwander, F., Eibach, R., Fechter, I., Hausmann, L., Zyprian, E., and Töpfer, R. (2012). Rpv10: a new locus from the Asian Vitis gene pool for pyramiding downy mildew resistance loci in grapevine. *Theor Appl Genet* 124, pp. 163–176. doi: 10.1007/s00122-011-1695-4
- Shi, X., Cao, S., Wang, X., Huang, S., Wang, Y., Liu, Z., et al. (2023). The complete reference genome for grapevine (Vitis vinifera L.) genetics and breeding. *Hortic. Res.* 10, uhad061. doi: 10.1093/hr/uhad061
- Suehiro, Y., Mochida, K., Itamura, H., and Esumi, T. (2014). Skin browning and expression of PPO, STS, and CHS genes in the grape berries of 'Shine muscat'. *J. Japanese Soc. Hortic. Sci.* 83, 122–132. doi: 10.2503/jjshs1.CH-095
- Szmania, C., Waber, J., Bogs, J., and Fischer, U. (2023). Sensory and aroma impact of mitigation strategies against sunburn in Riesling. *Oeno One* 57, 127–140. doi: 10.20870/oeno-one.2023.57.3.7287
- Tafel, M., and Szolnoki, G. (2020). Estimating the economic impact of tourism in German wine regions. *J. Tourism Res.* 22, 788–799. doi: 10.1002/jtr.2380
- Töpfer, R., and Trapp, O. (2022). A cool climate perspective on grapevine breeding: climate change and sustainability are driving forces for changing varieties in a traditional market. *Theor. Appl. Genet.* 135, 3947–3960. doi: 10.1007/s00122-022-04077-0
- van Leeuwen, C., Sgubin, G., Bois, B., Ollat, N., Swingedouw, D., Zito, S., et al. (2024). Climate change impacts and adaptations of wine production. *Nat. Rev. Earth Environ.* 5, 258–275. doi: 10.1038/s43017-024-00521-5
- Velt. in press. *GREAT (Grape expression ATLAS)*. Available online at: <https://great.colmar.inrae.fr/> (Accessed October 16, 2024). in prep.
- Velt, A., Frommer, B., Blanc, S., Holtgräwe, D., Duchêne, É., Dumas, V., et al. (2023). An improved reference of the grapevine genome reasserts the origin of the PN40024 highly homozygous genotype. *G3 (Bethesda)* 13. doi: 10.1093/g3journal/jkad067
- Waber, J., Bogs, J., Müller, K., and Friedel, M. (2023). Sonnenschutz: sonnenbrandschäden. *Der Deutsche Weinbau Neustadt* 24–27.
- Wang, X., Niu, Y., and Zheng, Y. (2021). Multiple functions of MYB transcription factors in abiotic stress responses. *Int. J. Mol. Sci.* 22, 6125. doi: 10.3390/ijms22116125
- Webb, L., Whiting, J., Watt, A., Hill, T., Wigg, F., Dunn, G., et al. (2010). Managing Grapevines through Severe Heat: A Survey of Growers after the 2009 Summer Heatwave in South-eastern Australia. *J. Wine Res.* 21, 147–165. doi: 10.1080/09571264.2010.530106
- Welter, L. C., Göktürk-Baydar, N., Akkurt, M., Maul, E., Eibach, R., Töpfer, R., et al. (2007). Genetic mapping and localization of quantitative trait loci affecting fungal disease resistance and leaf morphology in grapevine (Vitis vinifera L.). In *Mol Breeding* 20 (4), pp. 359–374.
- Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (Switzerland: Springer).
- Zha, Q., Xi, X., He, Y., and Jiang, A. (2020). Transcriptomic analysis of the leaves of two grapevine cultivars under high-temperature stress. *Sci. Hortic.* 265, 109265. doi: 10.1016/j.scienta.2020.109265
- Zhang, C., Dai, Z., Ferrier, T., Orduña, L., Santiago, A., Peris, A., et al. (2023). MYB24 orchestrates terpene and flavonol metabolism as light responses to anthocyanin depletion in variegated grape berries. *Plant Cell* 35, 4238–4265. doi: 10.1093/plcell/koad228
- Zhang, C., Navarro-Paya, D., and Matus, J. T. (2024). "Heat-stress responses regulated via a MYB24-MYC2 complex," in *Open GPB 2024 (IVES Conference Series)*.
- Zhang, M., Zhang, P., Lu, S., Ou-Yang, Q., Zhu-Ge, Y., Tian, R., et al. (2021). Comparative analysis of cuticular wax in various grape cultivars during berry development and after storage. *Front. Nutr.* 8. doi: 10.3389/fnut.2021.817796
- Zendler, D., Schneider, P., Töpfer, R., and Zyprian, E. (2017). Fine mapping of Ren3 reveals two loci mediating hypersensitive response against Erysiphe necator in grapevine. In *Euphytica* 213 (3). doi: 10.1007/s10681-017-1857-9
- Zschokke (1931). Beschädigung von Weinreben durch Sonnenbrand und Austrocknen. *Zeitschrift Für Pflanzenkrankheiten (Pflanzenpathologie) Und Pflanzenschutz*, 41(5), 240–251. Available at: <https://www.jstor.org/stable/43525981>.
- >Zyprian, E., Eibach, R., Trapp, O., Schwander, F., and Töpfer, R. (2018). Grapevine breeding under climate change: Applicability of a molecular marker linked to véraison. *Vitis: J. Grapevine Res.* 57, 119–123. doi: 10.5073/vitis.2018.57.119-123
- Zyprian, E., Ochßner, I., Schwander, F., Šimon, S., Hausmann, L., Bonow-Rex, M., et al. (2016). Quantitative trait loci affecting pathogen resistance and ripening of grapevines. *Mol. Genet. Genomics* 291(4), 1573–1594. doi: 10.1007/s00438-016-1200-5

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

