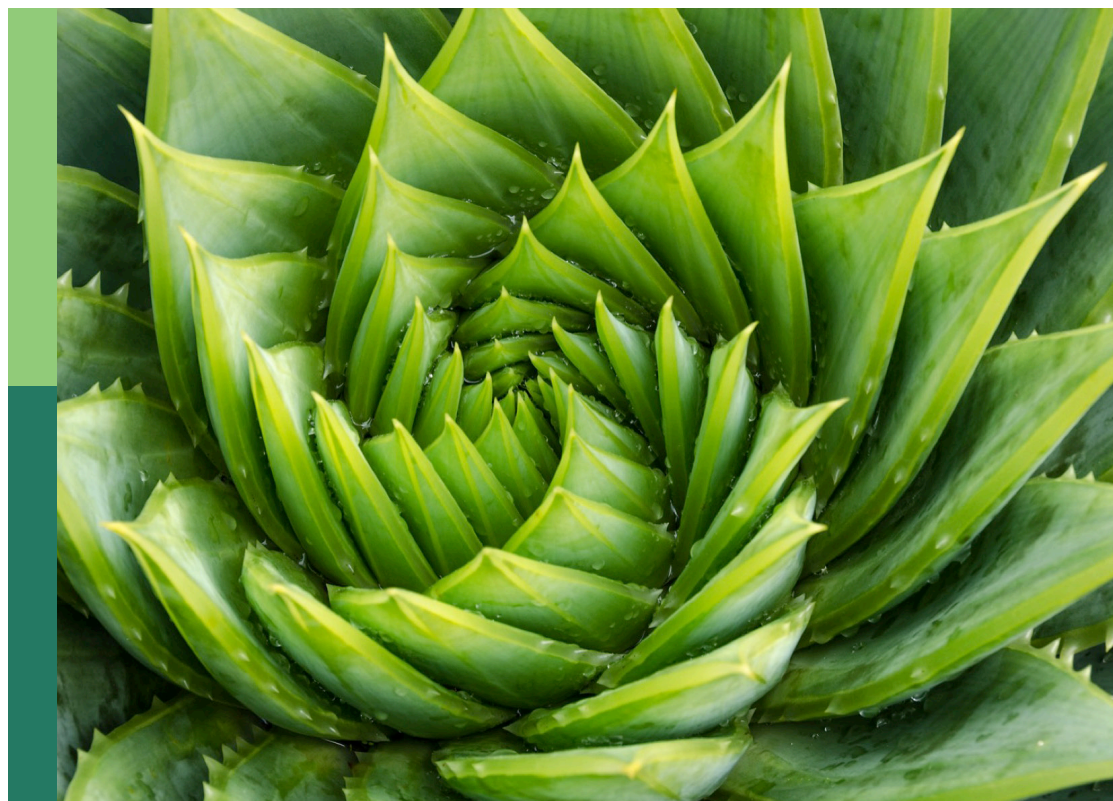


Insights in functional and applied plant genomics 2023

Edited by
Huihui Li

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 978-2-8325-6356-4
DOI 10.3389/978-2-8325-6356-4

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

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

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

Insights in functional and applied plant genomics: 2023

Topic editor

Huihui Li — Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, China

Citation

Li, H., ed. (2025). *Insights in functional and applied plant genomics: 2023*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6356-4

Table of contents

- 05 **Editorial: Insights in functional and applied plant genomics: 2023**
Huihui Li
- 08 **Advancing understanding of *Ficus carica*: a comprehensive genomic analysis reveals evolutionary patterns and metabolic pathway insights**
Yuting Bao, Miaohua He, Chenji Zhang, Sirong Jiang, Long Zhao, Zhengwen Ye, Qian Sun, Zhiqiang Xia and Meiling Zou
- 22 **Deep learning methods improve genomic prediction of wheat breeding**
Abelardo Montesinos-López, Leonardo Crespo-Herrera, Susanna Dreisigacker, Guillermo Gerard, Paolo Vitale, Carolina Saint Pierre, Velu Govindan, Zerihun Tadesse Tareegn, Moisés Chavira Flores, Paulino Pérez-Rodríguez, Sofía Ramos-Pulido, Morten Lillemo, Huihui Li, Osval A. Montesinos-López and Jose Crossa
- 37 **Feature engineering of environmental covariates improves plant genomic-enabled prediction**
Osval A. Montesinos-López, Leonardo Crespo-Herrera, Carolina Saint Pierre, Bernabe Cano-Paez, Gloria Isabel Huerta-Prado, Brandon Alejandro Mosqueda-González, Sofia Ramos-Pulido, Guillermo Gerard, Khalid Alnowibet, Roberto Fritsche-Neto, Abelardo Montesinos-López and José Crossa
- 68 **Genome-wide profiling of WRKY genes involved in flavonoid biosynthesis in *Erigeron breviscapus***
Wanling Song, Shuangyan Zhang, Qi Li, Guisheng Xiang, Yan Zhao, Fan Wei, Guanghui Zhang, Shengchao Yang and Bing Hao
- 84 **Chromosome-level assembly of *Lindenbergia philippensis* and comparative genomic analyses shed light on genome evolution in Lamiales**
Bao-Zheng Chen, Da-Wei Li, Kai-Yong Luo, Song-Tao Jiu, Xiao Dong, Wei-Bin Wang, Xu-Zhen Li, Ting-Ting Hao, Ya-Hui Lei, Da-Zhong Guo, Xu-Tao Liu, Sheng-Chang Duan, Yi-Fan Zhu, Wei Chen, Yang Dong and Wen-Bin Yu
- 101 **The pleiotropic functions of GOLDEN2-LIKE transcription factors in plants**
Mengyi Zheng, Xinyu Wang, Jie Luo, Bojun Ma, Dayong Li and Xifeng Chen
- 113 **Genome-wide analysis of the *WOX* gene family and function exploration of *RhWOX331* in rose (*R. 'The Fairy'*)**
Lian Duan, Zhihui Hou, Wuhua Zhang, Shuang Liang, Minge Huangfu, Jinzhu Zhang, Tao Yang, Jie Dong and Daidi Che

127 Advancing crop improvement through GWAS and beyond in mung bean

Syed Riaz Ahmed, Muhammad Jawad Asghar, Amjad Hameed, Maria Ghaffar and Muhammad Shahid

155 Screening and functional characterization of salt-tolerant NAC gene family members in *Medicago sativa* L

Zhiguang Li, Qianqian Yu, Yue Ma, Fuhong Miao, Lichao Ma, Shuo Li, Huajie Zhang, Zeng-Yu Wang, Guofeng Yang and Kunlong Su



OPEN ACCESS

EDITED AND REVIEWED BY
Peng Wang,
Jiangsu Province and Chinese Academy of
Sciences, China

*CORRESPONDENCE

Huihui Li
✉ lihuihui@caas.cn

RECEIVED 21 April 2025

ACCEPTED 23 April 2025

PUBLISHED 08 May 2025

CITATION

Li H (2025) Editorial: Insights in functional
and applied plant genomics: 2023.
Front. Plant Sci. 16:1615289.
doi: 10.3389/fpls.2025.1615289

COPYRIGHT

© 2025 Li. 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: Insights in functional and applied plant genomics: 2023

Huihui Li^{1,2*}

¹State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese
Academy of Agricultural Sciences (CAAS), CIMMYT-China Office, Beijing, China, ²Nanfan Research
Institute, Chinese Academy of Agricultural Sciences (CAAS), Sanya, Hainan, China

KEYWORDS

plant genomics, transcription factors (TFs), genome assembly, genomic prediction,
deep learning

Editorial on the Research Topic

Insights in functional and applied plant genomics: 2023

Recent advances in plant genomics have profoundly deepened our understanding of plant biology and accelerated efforts to address key agricultural challenges. The increasing availability of high-quality reference genome sequences for nearly all major crop species (Xie et al., 2024), alongside the widespread adoption of multi-omics platforms and cutting-edge gene-editing technologies (Rönspies et al., 2021), is enabling unprecedented insights into gene functions underlying critical phenotypic traits. These developments are paving the way for innovative strategies in crop improvement. However, the full potential of these datasets remains constrained by limitations in computational tools—particularly the lack of efficient algorithms for extracting biologically meaningful insights from large-scale datasets and the scarcity of robust machine learning models for accurate phenotype prediction in genomic selection (Farooq et al., 2024). Addressing these challenges will require parallel advances in biological and bioinformatics research to further accelerate the genetic improvement of crop plants.

Over the past decade, the field of crop genomics has made remarkable strides. As one of the most dynamic areas within plant sciences, it holds immense promise for ensuring food security and advancing sustainable agricultural development. This Research Topic, *Insights in Functional and Applied Plant Genomics: 2023*, was dedicated to exploring novel insights, emerging methodologies, ongoing challenges, and future directions in functional and applied plant genomics. The Research Topic features nine manuscripts, including seven original research articles, one review, and one systematic review. These contributions collectively span a wide range of plant systems, covering recent discoveries in both major crops like wheat (*Triticum aestivum* L.) and underrepresented species such as fig, rose, *Lindenbergia philippensis*, mungbean, and alfalfa.

A central theme in this Research Topic is the role of transcription factors (TFs) in regulating gene expression and phenotypic traits. Several articles explore the multifaceted

functions of TFs across diverse plant species. For instance, Zheng et al. provide a comprehensive review of the GOLDEN2-LIKE (GLK) TFs, which are functionally redundant nuclear regulators in the GARP subfamily of MYB transcription factors. GLKs are known to govern genes involved in photosynthesis and chloroplast biogenesis. Previous studies have shown that GLK knockout mutants display abnormal chloroplast structures without a complete loss of chloroplast formation. Zheng et al. synthesized current knowledge on the pleiotropic roles of GLKs, underscoring their broader functional relevance in plant biology.

Similarly, WRKY transcription factors—another major family—are explored for their key roles in plant development, stress responses, and secondary metabolite biosynthesis. In the medicinal plant *Erigeron breviscapus*, a species valued for its flavonoid content and therapeutic use in cardiovascular and cerebrovascular treatments, Song et al. identified 75 WRKY TFs through genome-wide analysis. Notably, 74 of these responded to exogenous treatments with abscisic acid and salicylic acid, and several were upregulated following gibberellin 3 (GA₃) application. Functional analysis revealed that many of these TFs were involved in flavonoid biosynthesis pathways. This study advances our understanding of WRKY-mediated regulation of secondary metabolism and offers promising avenues for breeding *E. breviscapus* cultivars with enhanced scutellarin content.

Alfalfa (*Medicago sativa* L.), one of the most widely cultivated forage crops, is renowned for its high tolerance to soil salinity. Li et al. conducted a genome-wide analysis of the *Zhongmu 1* cultivar and identified 114 members of the NAC transcription factor family, classifying them into 13 subgroups. Among these, subfamily V was found to play a potential role in salinity stress responses. Expression profiling revealed that *MsNAC40* plays a critical role in modulating salt stress. Functional validation through overexpression lines demonstrated significantly increased plant height, net photosynthetic rate, stomatal conductance, K⁺/Na⁺ ratio, and transpiration rate compared to controls, while leaf conductivity was significantly reduced.

WOX transcription factors, a plant-specific family, are also crucial in regulating development and stress responses. Duan et al. performed a comprehensive identification of 381 WOX genes in rose (*Rosa hybrida*), which were mapped across seven chromosomes. Transcriptome analysis revealed nine *RhWOX* genes with differential expression during root development, with three positively correlated with adventitious root formation. Notably, *RhWOX331* was shown to promote adventitious root primordium initiation. Overexpression of *RhWOX331* in *Arabidopsis thaliana* mitigated root growth inhibition by high concentrations of IBA and NPA, increased the number of lateral roots along the primary root, and enhanced overall plant height.

The assembly of new genome sequences from underutilized plant species provides critical insights into plant evolution and serves as a foundation for modern breeding approaches. Chen et al. assembled the genome of *L. philippensis*, an ornamental species collected from the Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences. The assembled genome was 407.46

Mb in size, with a level of completeness comparable to 15 other species within the Lamiales order, which comprises over 23,755 species across 24 families. This assembly contributes valuable information for understanding species diversification and genome evolution in Lamiales.

Another notable genome assembly was presented by Bao et al. for fig (*Ficus carica* L.), a nutritionally important horticultural crop. Cultivated for over 11,000 years in Southwest Asia and the Middle East, figs are known for their resilience to poor soils and harsh environmental conditions. The genome, spanning 366.34 Mb and assembled into 13 chromosomes, achieved a contig N50 length of 9.78 Mb. Comparative genomic analysis revealed that *F. carica* diverged from *F. microcarpa* approximately 2–3 million years ago, likely following a whole-genome duplication event. Additionally, allelic variation in the *CHS* gene in *F. carica* was shown to influence anthocyanin biosynthesis, contributing to differences in fruit color.

Advancements in genomic prediction rely heavily on the development and refinement of machine learning (ML) models. Montesinos-López et al. evaluated the performance of a deep learning (DL) model against the widely used genomic best linear unbiased prediction (GBLUP) method using a large wheat dataset. The DL model consistently outperformed GBLUP in predictive accuracy for two traits across a five-fold cross-validation framework. In a related study, Montesinos-López et al. demonstrated that integrating environmental covariates into genomic prediction models significantly enhances accuracy by reducing prediction errors. They validated this approach in maize and rice, emphasizing the importance of incorporating environmental data in genomic prediction. Nonetheless, further research is needed to establish robust feature engineering frameworks for effectively integrating environmental variables.

Mungbean (*Vigna radiata* L.), an important legume crop widely cultivated in South Asia and arid regions of southern Europe, has gained attention for its adaptability and nutritional value. Ahmed et al. provided an in-depth review of genome-wide association studies (GWAS) in mungbean, emphasizing their role in uncovering the genetic basis of agronomic traits and enhancing crop productivity through molecular breeding.

In summary, this Research Topic highlights pivotal advances in functional and applied plant genomics, with particular emphasis on underrepresented and neglected crop species. The Research Topic offers valuable insights into genetic mechanisms, novel genome assemblies, and emerging computational approaches, underscoring their collective potential to drive future crop improvement efforts.

Author contributions

HL: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported

by the National Natural Science Foundation of China (32361143514), Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-CSIAF-202303).

Conflict of interest

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

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.

References

Farooq, M. A., Gao, S., Hassan, M. A., Huang, Z., Rasheed, A., Hearne, S., et al. (2024). Artificial intelligence in plant breeding. *Trends Genet.* 40, 891–908. doi: 10.1016/j.tig.2024.07.001

Rönspies, M., Dorn, A., Schindele, P., and Puchta, H. (2021). CRISPR–Cas-mediated chromosome engineering for crop improvement and synthetic biology. *Nat. Plants* 7, 566–573. doi: 10.1038/s41477-021-00910-4

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.

Xie, L., Gong, X., Yang, K., Huang, Y., Zhang, S., Shen, L., et al. (2024). Technology-enabled great leap in deciphering plant genomes. *Nat. Plants* 10, 551–566. doi: 10.1038/s41477-024-01655-6



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Tingting Guo,
Huazhong Agricultural University, China
Qing Ma,
Zhejiang Shuren University, China

*CORRESPONDENCE

Zhiqiang Xia

✉ zqxia@hainanu.edu.cn

Meiling Zou

✉ mlzou@hainanu.edu.cn

[†]These authors have contributed equally to this work

RECEIVED 21 September 2023

ACCEPTED 20 November 2023

PUBLISHED 07 December 2023

CITATION

Bao Y, He M, Zhang C, Jiang S, Zhao L, Ye Z, Sun Q, Xia Z and Zou M (2023) Advancing understanding of *Ficus carica*: a comprehensive genomic analysis reveals evolutionary patterns and metabolic pathway insights. *Front. Plant Sci.* 14:1298417. doi: 10.3389/fpls.2023.1298417

COPYRIGHT

© 2023 Bao, He, Zhang, Jiang, Zhao, Ye, Sun, Xia and Zou. 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.

Advancing understanding of *Ficus carica*: a comprehensive genomic analysis reveals evolutionary patterns and metabolic pathway insights

Yuting Bao^{1†}, Miaohua He^{1†}, Chenji Zhang^{1,2}, Sirong Jiang¹, Long Zhao^{1,3}, Zhengwen Ye⁴, Qian Sun^{1,5}, Zhiqiang Xia^{1*} and Meiling Zou^{1*}

¹Sanya Nanfan Research Institute of Hainan University, Hainan Yazhou Bay Seed Laboratory, Sanya, China, ²College of Agriculture, China Agricultural University, Beijing, China, ³Academy of Agriculture and Forestry Sciences, Qinghai University, Xining, Qinghai, China, ⁴Forestry and Fruit Research Institute, Shanghai Academy of Agricultural Sciences, Shanghai, China, ⁵College of Life Science and Technology, Guangxi University, Guangxi, China

Ficus carica L. (dioecious), the most significant commercial species in the genus *Ficus*, which has been cultivated for more than 11,000 years and was one of the first species to be domesticated. Herein, we reported the most comprehensive *F. carica* genome currently. The contig N50 of the Orphan *fig* was 9.78 Mb, and genome size was 366.34 Mb with 13 chromosomes. Based on the high-quality genome, we discovered that *F. carica* diverged from *Ficus microcarpa* ~34 MYA, and a WGD event took place about 2–3 MYA. Throughout the evolutionary history of *F. carica*, chromosomes 2, 8, and 10 had experienced chromosome recombination, while chromosome 3 saw a fusion and fission. It is worth proposing that the chromosome 9 experienced both inversion and translocation, which facilitated the emergence of the *F. carica* as a new species. And the selections of *F. carica* for the genes of recombination chromosomal fragment are compatible with their goal of domestication. In addition, we found that the *F. carica* has the *FhAG2* gene, but there are structural deletions and positional jumps. This gene is thought to replace the one needed for female common type *F. carica* to be pollinated. Subsequently, we conducted genomic, transcriptomic, and metabolomic analysis to demonstrate significant differences in the expression of *CHS* among different varieties of *F. carica*. The *CHS* playing an important role in the anthocyanin metabolism pathway of *F. carica*. Moreover, the *CHS* gene of *F. carica* has a different evolutionary trend compared to other *Ficus* species. These high-quality genome assembly, transcriptomic, and metabolomic resources further enrich *F. carica* genomics and provide insights for studying the chromosomes evolution, sexual system, and color characteristics of *Ficus*.

KEYWORDS

Ficus carica, chromosome evolution, genome, *FhAG2*, *CHS*

Introduction

Ficus carica L. (*Fig*), a member of the Moraceae family's genus *Ficus*, is a heterozygous species (Mori et al., 2017). The reason it is termed “*fig*” is that the tiny flowers that are concealed in the hypanthium are not visible from the outside, only the pseudo-fruit formed by the receptacle. Generally accepted to have originated in Southwest Asia and the Middle East, *F. carica* are among the earliest known domesticated species, having been grown over 11,000 years (Kislev et al., 2006; Simsek et al., 2020). This species can tolerate extreme environmental conditions and poor soils (Vangelisti et al., 2019), and offer significant nutritional and health benefits (Vinson et al., 2005; Solomon et al., 2006; Veberic et al., 2008). *F. carica* has garnered a lot of attention lately as a promising functional food and drug candidate with high pharmacological activity because of their remarkable flavor and a variety of bioactivities (Purnamasari et al., 2019; Ayuso et al., 2022).

F. carica has significant therapeutic qualities, and research has looked into the possible use of common *fig* in the treatment of COVID-19 infections (Hamed et al., 2023). In addition, *F. carica* has a significant commercial value and is a major crop in the majority of Mediterranean nations as well as the US. The production of *F. carica* is anticipated to exceed one million tons year, the fruit's consumption has increased globally and is predicted to continue growing in the years to come (Harzallah et al., 2016).

There are over 800 species in the genus *Ficus*, which is one of the largest genera in angiosperms. It was discovered that most *Ficus* species were diploid, having 13–14 chromosomes ($2n=26$, $2n=28$). Among them, *Ficus microcarpa* (*F. microcarpa*) has 13 chromosomes with an assembled genome size of 436 Mb and is monoecious, while *Ficus hispida* (*F. hispida*) has 14 chromosomes and is dioecious (Zhang X. et al., 2020). *Ficus erecta* (*F. erecta*), a wild relative of common *F. carica*, has a genome size of 331.6 Mb and a Contig N50 of 1.9 Mb (Shirasawa et al., 2020). The first reported *F. carica* genome sequence is the Japanese cultivar, Horaishi, with a total assembled genome length of 248 Mb, Contig N50 of 4.5 Kb, and an estimated size of 356 Mb. The total length of the assembled genome is approximately 30% shorter than the estimated size (Mori et al., 2017). Later, the genome of another Italian *fig*, “Dottato,” was also published, with a total length of 333 Mb and a Contig N50 of 823 Kb. And 80% of the assembled genome was allocated to 13 chromosomes (Usai et al., 2020). As the most commercially significant species in the *Ficus* genus (Mawa et al., 2013), *F. carica* requires the assembly of a more comprehensive genome. In addition, chromosomes, which contain crucial genetic information for eukaryotes, have experienced a variety of intricate alterations during the course of the lengthy evolutionary engineering of organisms. Genome-wide duplication events are the first type of alterations in chromosome number, their importance in speciation and the development of new species cannot be overlooked (Ruprecht et al., 2017) and repeated rounds of WGD events can periodically boost plant genetic diversity (Mandakova et al., 2010). Chromosome chance events in unique contexts are the second category, wherein the number of individual chromosomes is either increased or decreased. The offspring inherit this alteration steadily and with retention. Chromosome

rearrangement is one of the most interesting chromosomal occurrences. When DNA double strand breaks are being repaired, an unusual type of recombination called chromosome rearrangement takes place. Chromosome rearrangement consists of a variety of changes such as chromosome insertions, deletions or duplications, inversions, translocations, and transpositions. The term “chromosomal translocation” refers to the movement of chromosome segments from one chromosome to another, duplicate chromosome segments on distinct chromosomes are also thought to be the outcome of this process. Translocation and inversion are the two types of chromosomal recombination that can lead to secondary recombination of chromosomes and changes in chromosome structure (Schubert and Lysak, 2011), which can further modify the karyotype of the organism. These have the potential to alter chromosomal numbers, which could lead to the emergence of new species and the diversification of existing ones (Schubert and Lysak, 2011; Romanenko et al., 2019).

Caprifig, Smyrna, San Pedro, and common *fig* are the four types of *F. carica* that can be distinguished by their reproductive and pollination traits. *F. carica* trees are gynodioecious with two majors sex types: the caprifig and *fig* types. Though caprifigs are hermaphrodite plants with both male and female blooms, they solely function as male plants because they can only bear pollen and not edible fruit. However, female blooms continue to be essential for artificial feminization or wasp pollination of some *fig* species (Ikegami et al., 2013). More than half of *Ficus* plants display dioecy from a functional standpoint. Reportedly, the sexual orientation of *F. carica* is determined by the *RAN1* gene (Mori et al., 2017). However, studies have demonstrated that the *RAN1* gene does not exhibit clear gender or organ specificity in its expression. The *AGAMOUS* paralogous homologous gene *FhAG2* was shown to be the candidate gene accountable for male-specific gender identity in *hispida* (Zhang, X., et al., 2020). These investigations offer guidance and important data for the study of the genomes of *Ficus* plants, and they will be important for future investigations into the genes of dioecious and unisexual plants. Further, *F. carica* comes in a variety of colors, including yellow, green, red, purple. Researches have shown that the red peel of *F. carica* is mainly determined by the content of anthocyanins, while the yellow peel is mainly due to the high content of carotenoids, the green peel is mainly result from the high content of chlorophyll. *F. carica* peels contain four different types of anthocyanins: pelargonin-3-glucoside, cyanidin-3,5-diglucoside, cyanidin 3-glucoside, and cyanidin-3-rutinoside (Duenas et al., 2008; Treutter et al., 2010; Zhang H. et al., 2020). Previous researchers have examined the anthocyanin biosynthesis route. Essentially, 4-coumaric acid is produced by phenylalanine, and 4-coumaric acid CoA ligase (4CL) catalyzes the creation of 4-coumaric acid 4-coumaric CoA. 4-Enzymes involved in anthocyanin synthesis work with fumaric acid CoA and another precursor, malonyl CoA, to produce stable anthocyanins in the end (Castellarin et al., 2007; Czemplak et al., 2012; Zhang et al., 2014). The anthocyanidin biosynthesis pathways in plants have been extensively studied (Tanaka et al., 2008) and are associated with many genes and transcription factors. However, the ‘anthocyanin synthesis pathway’ related to the variations in the flesh color in the different varieties of *F. carica* has been rarely studied.

Currently, there is a vast and varied range of *F. carica* varieties, and scientific research on *F. carica* is continually growing. Genetic and breeding studies will have greater benefit from a more complete genome. Furthermore, no publications have been published on the transcriptome and secondary metabolome of different varieties, which is extremely important for *F. carica* genome mining and genetic improvement. In this study, Orphan fig was selected as the research material and third-generation long-segment nanopore sequencing was used to sequence the young and fresh leaves of Orphan fig. As a reference genome, the excellent Orphan genome was built. Joint analysis was performed using the acquired *F. carica* genome in conjunction with transcriptome and secondary metabolome analysis.

Materials and methods

Plant material and library construction

Utilizing improved Cetyltrimethylammonium Bromide (CTAB) method was used to extract lengthy DNA segments weighing more than 500 ng from the tender leaves of Orphan (A212) (Supplementary Figure 1). Afterwards, the purified library was sequenced using a nanopore sequencer (Oxford Nanopore Technologies, Oxford, UK). After tender fig leaves were fixed in formaldehyde, the cells were lysed, and samples were taken out to assess the quality. Following biotin labeling, blunt-end ligation, chromatin digestion with restriction enzymes, DNA extraction, and purification, Hi-C samples were made and their DNA quality examined. A standard library was built once the quality test was passed. The NovaSeq platform (Illumina, San Diego, CA, USA) was used for the sequencing. Fastp (version 0.23.0) (Chen et al., 2018) with default parameters was used to filter adaptor contamination and low-quality reads in order to get clean sequencing data.

Genome assembly and quality assessment

Nanopore-derived reads were corrected using NextDenovo (<https://github.com/Nextomics/NextDenovo>) and then used as input for SMARTdenovo assembly (Istace et al., 2017). After the initial assembly, polishing was repeated with NextPolish (Hu et al., 2020). The valid end reads obtained based on the Hi-C data were used to assist with genome assembly. Using 3D DNA pipeline (<https://github.com/theaidenlab/3d-dna>), the contigs were divided into subgroups and reassembled (Olga et al., 2017). In addition, a BUSCO (Benchmarking Universal Single-Copy Orthologs) (Seppey et al., 2019) assessment of the genome was performed to evaluate the entirety of the assembled genome.

Genome and TF annotation

Repeat sequences in the *F. carica* genome were identified based on self-BLAST (<https://github.com/Dfam-consortium/>

RepeatModeler) using the RepeatModeler (version 1.0.10) (<https://github.com/Dfam-consortium/RepeatModeler>) (Flynn et al., 2020). RepeatMasker (version 4.0.7) (<http://www.repeatmasker.org>) cross-matching was used to search further for known repeats. A pipeline integrating *de novo* gene prediction and RNA-seq gene model was used to predict the protein-encoding genes. For *de novo* gene prediction, Augustus (version 3.0.2) (Stanke et al., 2006) and SNAP (<https://github.com/KorfLab/SNAP>) were run with default parameters. For RNA-seq-based prediction, RNA-seq reads were screened from the pooled tissue samples to eliminate the adapters and trimmed to remove low-quality bases. The processed reads were then aligned with the reference genome.

Construction of evolutionary tree and estimation of evolution rate

Homologous gene families were identified in the genomes of *Ficus carica*, *Ficus hispida*, *Ficus microcarpa*, *Morus alba*, *Cannabis sativa*, *Ziziphus jujuba*, *Arabidopsis thaliana*, *Carica papaya*, *Citrus sinensis*, *Manihot esculenta*, *Vitis vinifera*. To construct protein gene sets of multiple species, the encoded protein sequences were obtained from the genomic data of the species mentioned. OrthoFinder (version 2.2.6) (Emms and Kelly, 2019) was employed to cluster the selected protein sequences and identify orthologous genes by screening for genes with low-copy numbers. Single-copy, homologous genes were identified from the collection and used to construct an evolutionary tree (Price et al., 2010). The evolutionary tree was converted to a time tree using r8s Calibrate Time of the Timetree database (<http://www.timetree.org/>) (Kumar et al., 2017). CAFÉ (version 4.1) (De et al., 2006) was employed to analyze the expansion and contraction of gene families based on the chronogram of the 11 species.

Collinearity and Ks analysis

MCSanX (Wang et al., 2012) set to default parameters was used to identify the collinear genes, and proteins were used to screen the genomes of species to obtain the best matching pair. Each aligned block represented an orthologous pair derived from a common ancestor. Ks (synonymous substitutions per synonymous site) values for the homologs within the collinear block were determined using PAML (version 4.5) (Yang, 2007). The median Ks value was regarded as the representative of the collinear block. The hypothetical whole-genome replication and the putative whole-genome duplication (WGD) events in *F. carica* were identified by plotting the values of all gene pairs. The formula $t = Ks/2r$ representing the neutral substitution rate was used to estimate the replication and differentiation times between *F. carica* and other species. The neutral substitution rate used in this study was 8.12×10^9 . Calculation of ka/ks was performed using the KaKs_Calculator (<http://evolution.genomics.org.cn/software.htm>) (Zhang et al., 2006) software. when Ka is equal to Ks ($ka/ks = 1$), indicating a neutral mutation; when Ka is less than Ks ($ka/ks < 1$), it indicates a

negative (purification) selection; when K_a exceeds K_s ($k_a/k_s > 1$), it indicates a positive (diversification) selection.

Analysis of the secondary metabolome

The samples from four varieties of *F. carica*, F1: “Orphan,” F2: “Balaonai,” F3: “Violette Solise,” and F4: “Bpjhon,” each with a different fruit flesh color were collected. The secondary metabolites from each variety were extracted in triplicate. Ultra-high-performance liquid chromatography was the primary analytical system used, and the data obtained was scrutinized by the Analyst 1.6.3 software. Metabolites with a fold change ≥ 2 or ≤ 0.5 , P -value < 0.05 , and item variable importance ≥ 1 were considered statistically significant. Using the KEGG compound database, the metabolites identified were annotated to the KEGG pathway database (Kanehisa et al., 2017).

Transcriptome sequencing

The fruits of F1, F2, F3, and F4 were harvested in three biological repetitions and immediately frozen in liquid N_2 . RNA seq analysis included RNA isolation, library construction, and sequencing for gene prediction. Raw data was trimmed to eliminate the adapters and improve the quality. The reads < 100 bp long were discarded. TopHat2 (version 2.0.4) (Kim et al., 2013) was used for mapping the clean reads to the genome under default parameters. The transcript was assembled using Cufflinks (version 2.2.1) (Trapnell et al., 2012). The gene expression levels were measured using transcriptional fragments plotted from Cufflinks per kilobase bases per million fragments, and the differentially expressed genes (DEGs) were determined with DESeq2 (Varet et al., 2016). The expression data for different breeds were centralized, normalized, and then clustered using K-means to analyze the differential gene expression patterns. False discovery rates were used for adjusting P value. The genes with statistically significantly different expression levels, i.e., $|\log_2(\text{fold change})| \geq 1$ and adjusted P values < 0.05 , were identified as DEGs and annotated using the GO enrichment and KEGG pathways.

Functional gene analysis

Ten gene families in the *F. carica* genome, including *PAL*, *C4H*, *4CL*, *CHS*, *CHI*, *F3H*, *F3'H*, *DFR*, *ANS*, and *UGT*, were detected from the HMM domain model and using BLASTP (version 2.2.3.1) by studying the pathways involved in the regulation of fruit color, especially the “anthocyanin synthesis pathway.” *Fig* genome was screened for identification by employing the HMMER (version 3.0) software. Then, conserved domains were confirmed in all the protein sequences, while those with incomplete domains were excluded. The Pfam database (El-Gebali et al., 2019) was used to predict the domains of these protein homologs, and the genes that encoded proteins with identical domains were considered homologs.

Results

Sequencing and assembly of the genome

Using K -mer analysis, the size of the *F. carica* genome was estimated to be 356 Mb. In total, 3,301,024 reads amounting to 42 Gb were generated, with a sequencing depth of $\sim 100\times$ and an average read length of 12.78 Kb. A high-throughput chromosome concept capture (Hi-C) library of the genome of “Orphan” was constructed to enhance the quality of the assembly and mount the contigs on chromosomes, which resulted in 64.26 GB of Hi-C paired ends at $140\times$ (Supplementary Table 1).

After genome assembly, polishing, and elimination of redundancy, the final size of the genome was 373.72 Mb, and that of the contig N50 was 9.78 Mb (Supplementary Table 2). The Hi-C heatmap was first examined, and a diagonal pattern of high link frequencies was observed in the individual pseudochromosomes, indicating increased interactions between adjacent regions (Supplementary Figure 2). Duplicate deletions, classification, and quality assessment were performed on Hi-C-Pro, and only the mapped, valid reads were used for Hi-C. As a result, a sequence 366.34 Mb in length was allocated to 13 chromosomes, accounting for 98.02% of the total length, with the number of corresponding contig cut bins obtained being 3,906 (Supplementary Table 3). The completeness and accuracy of the assembled genome were assessed using BUSCO, and over 96.2% of the BUSCO-derived assessments were located in the assembled genome (Supplementary Figure 3, Supplementary Table 4) (Table 1).

Genome annotation

The *de novo* predicted genomic data were collected from the young leaves of *F. carica*, and the transcriptome data from the four samples of *F. carica*. Comparison between *F. carica* and *F. microcarpa* and *F. hispida*.

Collinearity analysis of *a*, the results show that *F. microcarpa* and *F. hispida* possessed 29,402 and 27,210 genes, respectively. In total, 29,783 protein-coding genes were identified within the *F. carica* genome, of which 29,039 were mapped to specific chromosomal loci, accounting for 97.5% of the total genome, with an average gene length of 3,111 bp, and a coding sequence length of 32,006,639 bp penetrance (exon). The mean GC content was 34.13%, higher than that reported in a previous study (33.38%). A total of 812,147 repeat sequences were identified in the assembled genome, accounting for 47.92% of the total genome, which too was higher than the value of 20.9% reported previously. Of these repeats, LTRs accounted for 10.11% (most abundant) and transposons for 3.26% (Supplementary Table 5) (Table 1).

Evolution, genome-wide replication, and species collinearity in *F. carica*

Collinearity analysis of the *F. carica*, *F. hispida*, and *F. microcarpa* genomes suggested that a total of 9,926 single-copy,

homologous gene pairs between *F. carica* and *F. microcarpa*; and 9,606 between *F. carica* and *F. hispida* were identified. Analysis of the homologous genes demonstrated a close evolutionary relationship between these three species. *F. hispida* had a substantial congruity with *F. carica*; chromosome 14 and chromosome 2 of *F. hispida* had a remarkable conformity with chromosome 1 of *F. carica*. In addition, 19,967 and 19,825 collinear genes were identified between *F. carica* and *F. hispida*, *F. microcarpa*, respectively, indicating that 67% and 66% of the *F. carica* genome was collinear with those of the respective species.

These results further indicated that the ancestors of *F. carica* may have undergone chromosomal fusions or divergences (Figure 1A, B).

A phylogenetic tree was constructed using r8s, which showed that *F. carica* was closely related to *Z. jujuba* (Rhamnaceae), *C. sativa* and *M. alba* (Moraceae). The divergence of *F. carica* and *F. hispida* from *F. microcarpa* occurred approximately 40 million years ago. *F. carica* then diverged from *F. hispida* about 34 million years ago. The expansion and contraction of gene families are crucial characteristic features of species selective evolution. Analysis

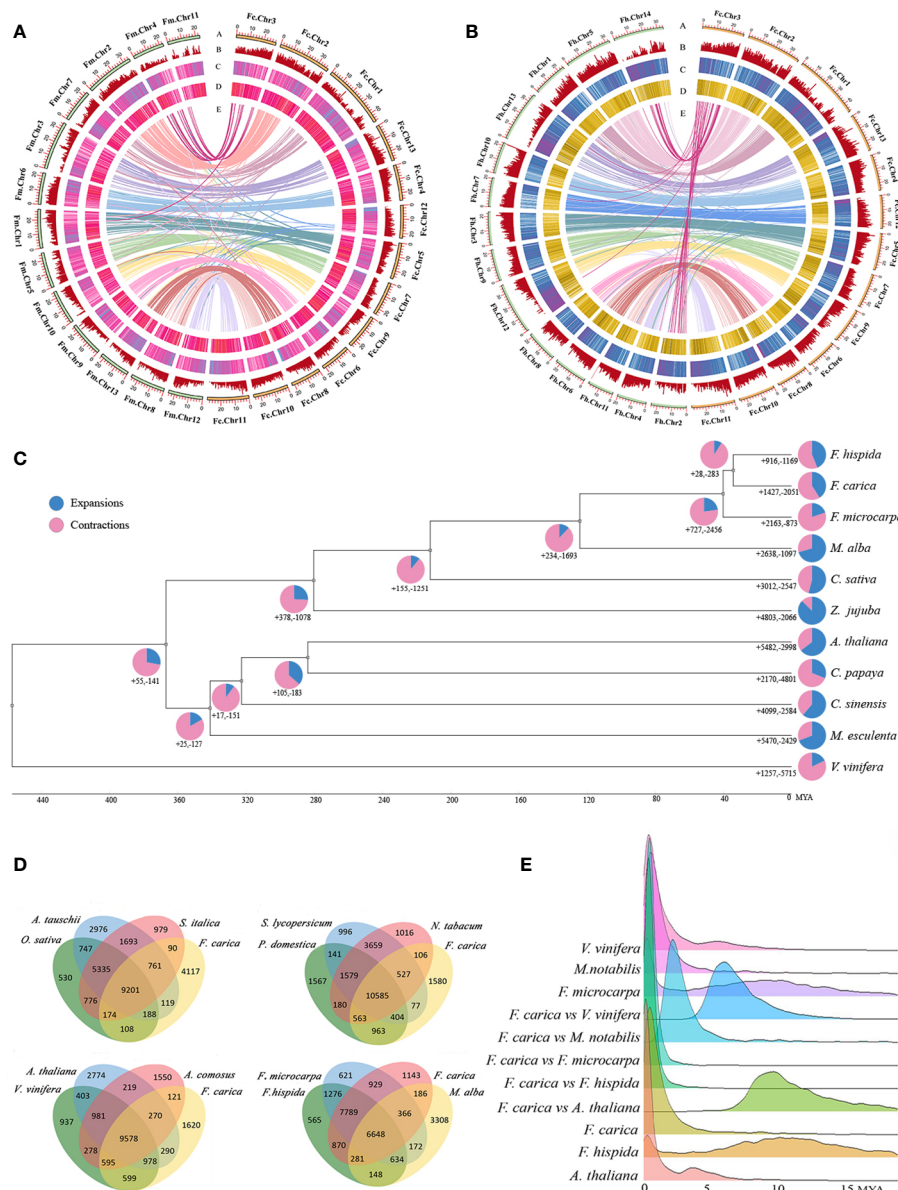


FIGURE 1

Phylogenomics and genomic evolution. (A) Genomic features of *Ficus carica* and *Ficus microcarpa*. (B) Genomic features of *Ficus carica* and *Ficus hispida*. The Circos plot of the multidimensional topography depicted from the outermost to innermost, (A–E) chromosome karyotypes, gene density, LTR TEs, DNA TEs, and synteny between the two genomes. (C) Phylogenetic tree, divergence times, and expansion and contraction of gene families in 10 species and *Ficus carica*. Pie charts indicate the proportion of gene families that underwent expansion and contraction. (D) Venn diagram showing the shared and unique gene families among *Ficus carica*, *Agelios tauschii*, *Oryza sativa*, *Setaria italica*, *Solanum lycopersicum*, *Prunus domestica*, *Nicotiana tabacum*, *Arabidopsis thaliana*, *Vitis vinifera*, *Ananas comosus*, *Ficus microcarpa*, *Ficus hispida*, *Morus alba*. (E) Ks distribution of *Vitis vinifera*, *Morus notabilis*, *Ficus microcarpa*, *Ficus carica*, *Ficus hispida*, and *Arabidopsis thaliana*, between *Ficus carica* and *Vitis vinifera*, *Morus notabilis*, *Ficus microcarpa*, *Ficus hispida*, and *Arabidopsis thaliana*.

showed that *F. carica*, *F. microcarpa*, and *F. hispida* acquired new genes and gene families during evolution. However, during the evolution of each species, independent gene families were acquired and lost to varying degrees. *F. hispida* and *F. carica* underwent expansion of 727 gene families and contraction of 2,456 gene families after differentiation from *F. microcarpa*. While in *F. carica*, 1,427 gene families were expanded, and 2,051 were contracted. In the *F. hispida* evolution node, 916 gene families were expanded, and 1,169 were contracted (Figure 1C).

Comparative analyses of the gene families revealed 9,201 gene families common to *O. sativa*, *S. italica*, and *A. tauschii*, 10,535 to *S. lycopersicum*, *P. domestica*, and *N. tabacum*, 9,578 to *V. vinifera*, *A. thaliana*, and *A. comosus* and 6,648 to *F. microcarpa*, *F. hispida*, and *M. alba*. Compared with *F. microcarpa* and *F. hispida*, 6,199 gene families were unique to *F. carica* (Figure 1D, Supplementary Table 6). GO analysis of these families revealed that ‘transporter activity’ and ‘transcription regulator activity’ were significantly enriched (Supplementary Figure 4). The protein sequences of *F. carica*, *F. hispida*, *F. microcarpa*, and *A. thaliana*, were compared using Blastp. The number of homologous genes identified was 2,943 between *F. carica* and *F. hispida*; 2,079 between *F. carica* and *F. microcarpa*; and 4,295 between *F. carica* and *A. thaliana*. There were 8,142 homologous genes in *F. carica*, 9,849 in *F. microcarpa*, and 6,272 in *F. hispida*. Based on the comparisons of the homologous genes, the time point when the *F. carica* underwent genome-wide replication was ascertained, and the synonymous substitution rate (KS) was calculated. The KS values of the homologous gene pairs between *F. hispida*, *F. microcarpa*, and *F. carica* were calculated to judge the time point of differentiation. The results obtained showed that *F. microcarpa* differentiated earlier than *F. carica* and *F. hispida*, while *F. hispida* and *F. carica* were closely related and differentiated in 5 million years, which was consistent with the results of the evolutionary time tree. Each peak of KS in the genome represents a genome-wide replication (WGD) event. However, the peak close to the vertical axis on the left results from repeats in the genome and is not an actual KS peak. Therefore, as shown in Figure 1E, *F. carica* must have undergone WGD events during evolution 2–3 million years.

Chromosomal evolutionary analysis of *F. carica* with *F. microcarpa* and *F. hispida*

Strong correspondences have been observed between the chromosomes of *F. carica*, *F. microcarpa*, and *F. hispida*. Several instances of chromosome fusion and breakage between the chromosomes of *F. carica* and the two *Ficus* species were discovered by further collinearity analysis. As *F. carica* differentiated from *F. microcarpa*, chromosomes 4 and 11 of *F. microcarpa* joined together to become chromosome 3 of *F. carica*. Moreover, the chromosome 1 of *F. microcarpa* splits to generate chromosomes 5 and 12 in the *F. carica*. In addition, chromosome fusion and breakage events between *F. carica* chromosome 3 and *F. hispida* chromosomes 2 and 14 occurred throughout the process of *F. carica* and *F. hispida* development. Chromosome variation has long been known to encourage the emergence of new species and

the diversification of existing ones. In the analysis of the covariation between *F. carica* and two *Ficus* species, it was found that after *F. carica* diverged from *F. microcarpa*, the inversion of the 3.59 Mb fragment at the end of *F. carica* chromosome 2, the chromosome duplication occurred on the 12.04 Mb fragment at the anterior end of chromosome 8 as well as on the 10.77 Mb fragment at the anterior end of chromosome 10, and furthermore the 12.44 Mb fragment at the anterior end of *F. carica* chromosome 9 also had a chromosomal translocation. In addition, chromosomal translocations and inversions occurred on the 12.44 Mb segment of the anterior end of *F. carica* chromosome 9. Coincidentally, after *F. carica* diverged from *F. hispida*, *F. carica* chromosome 2 was also inverted and duplications occurred on chromosomes 8 and 10, and similarly, chromosomal translocations and inversions occurred on the anterior end of *F. carica* chromosome 9 (Figure 2A, B). *F. carica* has a chromosome number of 13, whereas *F. hispida* has 14 chromosomes. This difference in chromosomal number can be attributed to the chromosomal recombination events previously mentioned.

GO enrichment was performed for genes with mutated segments on *F. carica* chromosome 2, chromosome 8 and 9, and chromosome 10. It was found that these genes were mainly enriched in ‘chalcone metabolic process’, ‘chalcone biosynthetic process’, and ‘chalcone synthase’. Chalcone synthase is the first enzyme in the synthesis pathway of plant flavonoids, which is not only closely related to plant fertility, but also plays an important role in plant resistance to pathogens. In addition to this, they were also enriched in ‘gametophyte development’, ‘sucrose synthase activity’, ‘sucrose biosynthetic process’, ‘sucrose metabolic process’ (Figure 2C). Compared to the two *Ficus* species, it is worth suggesting that *F. carica* trees are shorter in height and have edible fruits. The objective of their domestication is also reflected in the choice of *F. carica* for gene selection. This conclusion can be further demonstrated by looking at the selection pressure analysis of *F. carica* and *F. hispida*. Positive selection (Figure 2D) is mostly carried out by *F. carica* during the evolutionary process on genes associated with processes like ‘negative regulation of developmental growth’, ‘negative regulation of growth’, and ‘regulation of cellular biological process’ (Figure 2E).

Analysis of gene related to sex determination in *F. carica*

Both *F. carica* and *F. hispida* are dioecious *Ficus* species. In the dioecious *F. carica*, the protein-coding gene *Fh.AG2* unique to males was discovered. Sequence alignment revealed similar genes on chromosome 3 of *F. carica* and in the hermaphrodite *F. microcarpa*. The *Fh.AG2* gene jumped on the chromosomes of *F. carica* and *F. microcarpa*, according to a comparison of their positions on the chromosomes with *F. hispida* (Figure 3A). Alignment of gene protein domains showed two obvious deletions in the protein domain of *F. carica* compared with that of *F. hispida* and *F. microcarpa*. Furthermore, there were many protein-coding gene regions that *F. carica* and *F. hispida*, *F. microcarpa* shared (Supplementary Figure 5). The cis-acting

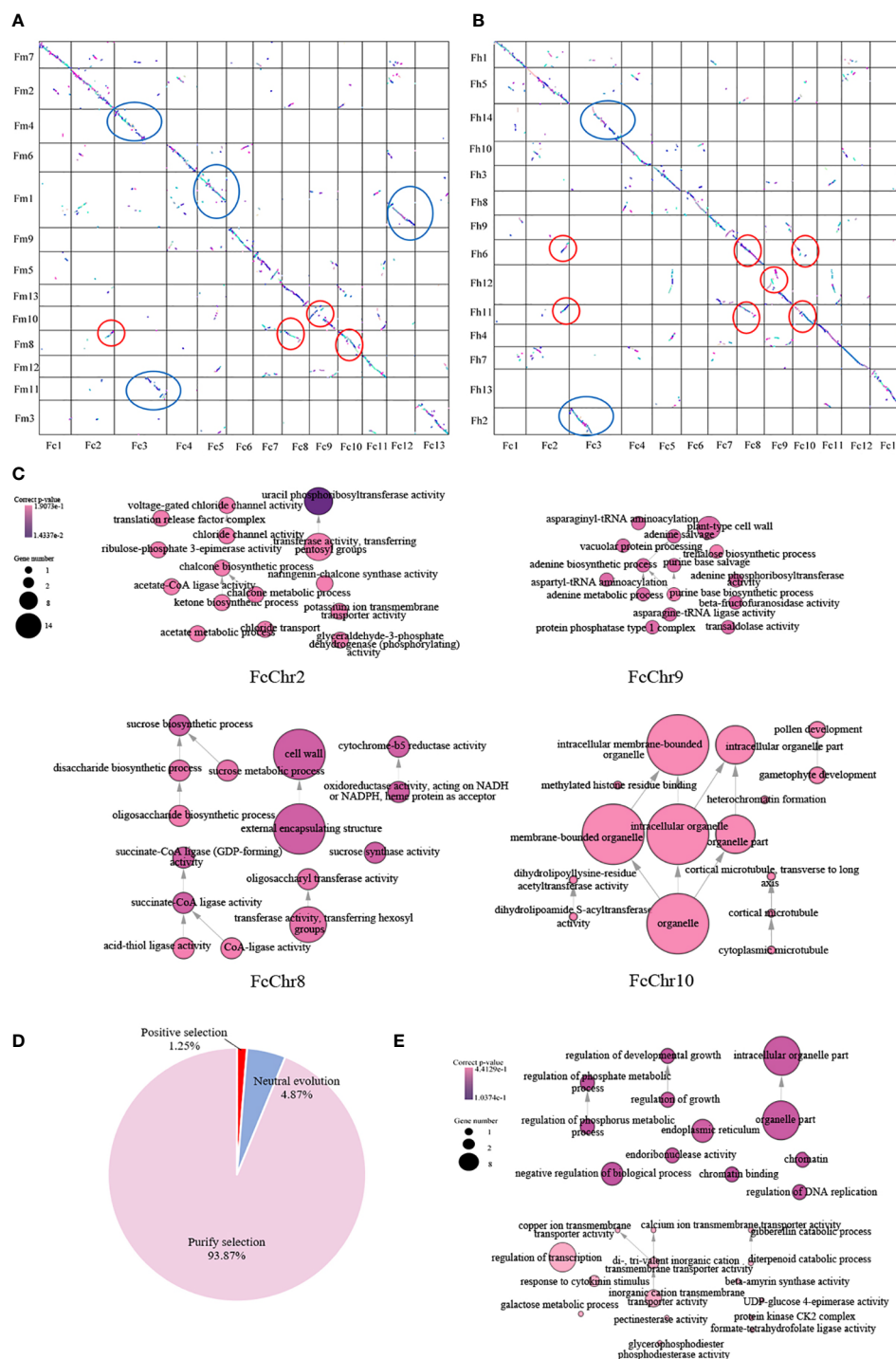


FIGURE 2
Chromosomal evolutionary analysis of *F. carica* with *F. microcarpa* and *F. hispida*. **(A)** Chromosomal collinearity of *Ficus carica* and *Ficus microcarpa*. **(B)** Chromosomal collinearity of *Ficus carica* and *Ficus hispida*. Chromosome fusion and fission are circled in blue ellipses, while chromosome recombination is outlined in red ellipses. Chromosome inversion and translocation are circled in red. **(C)** Functional enrichment of genes involved in recombination of *Ficus carica* chromosomes 2, 8, 9, and 10. **(D)** Analysis of selection pressure for *F. carica* and *F. hispida*, including positive selection ($k_a/k_s > 1$), neutral evolution ($k_a/k_s = 1$), purify selection ($k_a/k_s < 1$). **(E)** Functional enrichment of positive selection genes in *Ficus carica*.

elements in the promoters of this gene family were discovered by comparing the first 2000 bp of the AG gene promoters of these three *Ficus* plants. Additionally, it was discovered that the 400-2000 bp of *F. microcarpa* and the first 1600 bp of *F. hispida* shared comparable promoter structure. Compared with the other species of Italy *Ficus*,

the promoter in *F. hispida* was mainly a CAAT-box, and that in *F. carica* was mainly a CAAT-box and a TATA-box. The main functions of these promoters and the common *cis*-acting elements of the promoters and enhancers are transcription initiation around the core promoter element at -30, respectively (Figure 3B).

The AG gene structure comparison between *F. carica* and *F. hispida* and *F. microcarpa* revealed that there were clear deletions in the CDS structural domain of the *F. carica* genes. The two *Ficus* varieties and *F. carica* evolved relatively separately, according to the evolutionary tree built from the CDS sequences (Figure 3C). From the perspective of protein structure, the relationship between *F. hispida*, *F. carica*, and other dioecious plants was much stronger than that with *F. microcarpa* (Figure 3D). During evolution, the genes involved in sex determination in *F. hispida* and *F. carica* underwent a specific differentiation. The Ka/Ks ratio of *F. microcarpa* and *F. carica* was 1.04, indicating that the gene was affected by positive selection in these species.

Transcriptome sequencing, clustering, and functional enrichment

Transcriptome sequencing of four differently colored *F. carica* fruits was performed (Figure 4A). After filtering raw data, checking error rates, and determining the GC content, 45.98–63.77 million high-quality, 150 bp base raw data were obtained. The clean reads were then mapped to the genome, and more than 90% could be successfully aligned (Supplementary Table 7). From the transcriptomes of the four groups of samples, 2,582 DEGs (1,797 up-regulated and 785 down-regulated) were identified between F1 and F2, 3,445 DEGs (2,272 up-regulated and 1,173 down-regulated)

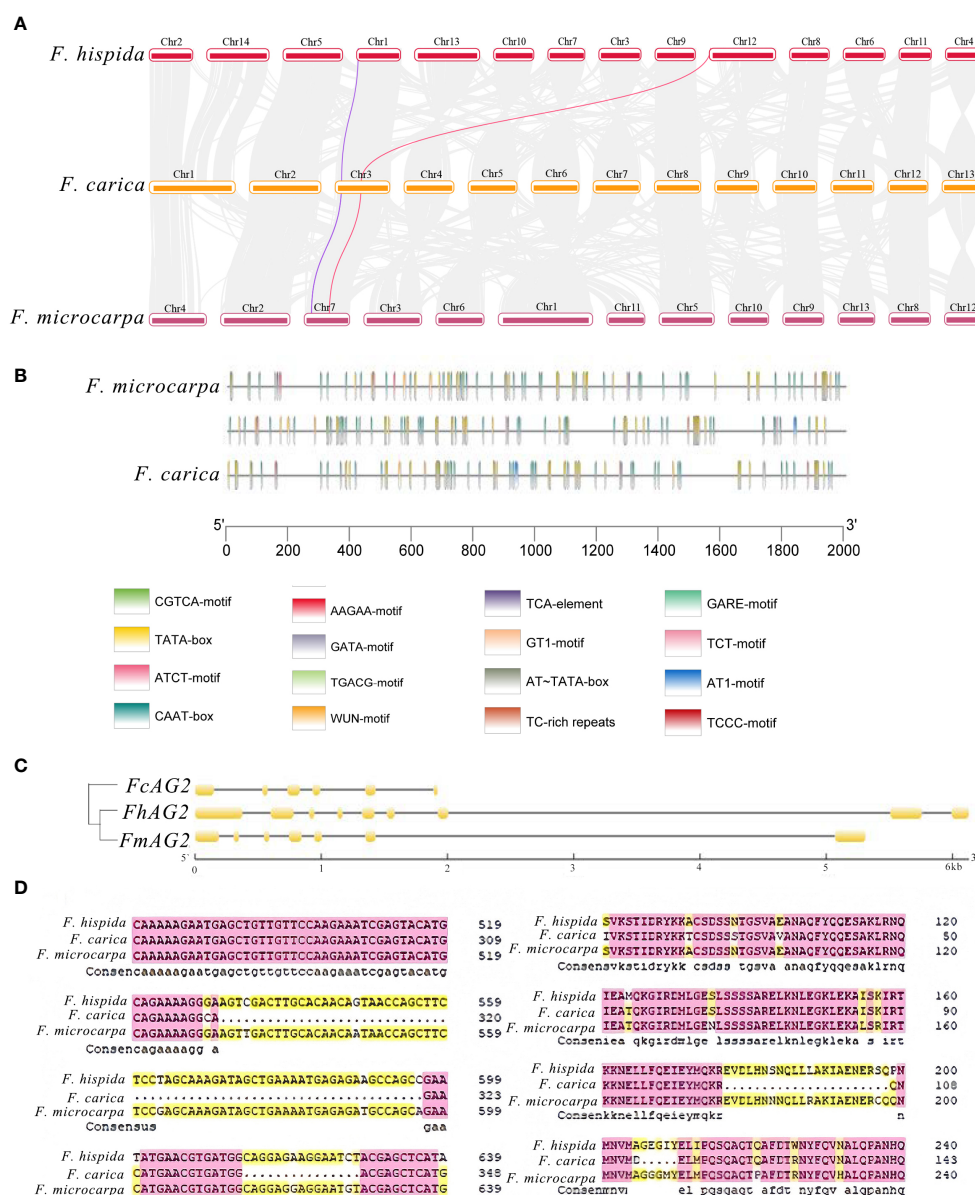


FIGURE 3

Analysis of Genes Related to Sex Determination in *F. carica*. (A) Collinearity analysis of the male and female sex-determining genes of *F. carica* and *F. hispida*, *F. microcarpa* (the purple line indicates *RAN1* and the red line, *FhAG2*). (B) CIS original analysis of the *F. carica*, *F. hispida*, and *F. microcarpa* promoters. (C) Structural analysis and the evolutionary tree of AG genes in the *F. carica*, *F. hispida*, and *F. microcarpa*. (D) Alignment of the *F. carica*, *F. hispida*, and *F. microcarpa* protein sequences.

between F1 and F3, and 3,062 DEGs (1,924 up-regulated and 1,138 down-regulated) between F1 and F4 (Supplementary Figure 6). In total, 980 DEGs and 47 differentially expressed TFs were also identified to be common to the four sets of data (Supplementary Table 8, Supplementary Figure 7B, C).

The expression patterns of the genes identified in the three stages of fruits were divided into ten subfamilies (Supplementary Figure 8). GO enrichment showed that F1 and F2 were significantly enriched in 'secondary metabolic processes', 'chloroplast thylakoid',

'plasma membrane fraction', 'plastid thylakoid', 'thylakoid', 'ADP binding', 'heme binding', 'monooxygenase activity', 'oxoquinene cyclase activity', 'tetrapyrrole binding', and 'UDP glucosyltransferase activity' (Supplementary Figure 9). For F1 and F3, 'bacterial defence responses', 'secondary metabolic processes', 'secondary metabolite biosynthesis', 'chloroplast thylakoid', 'components of the plasma membrane', 'parts of the plasma membrane', 'hydrolase activity acting on glycosidic bonds', 'hydrolase activity', 'hydrolysis of oxyglycosyl compounds',

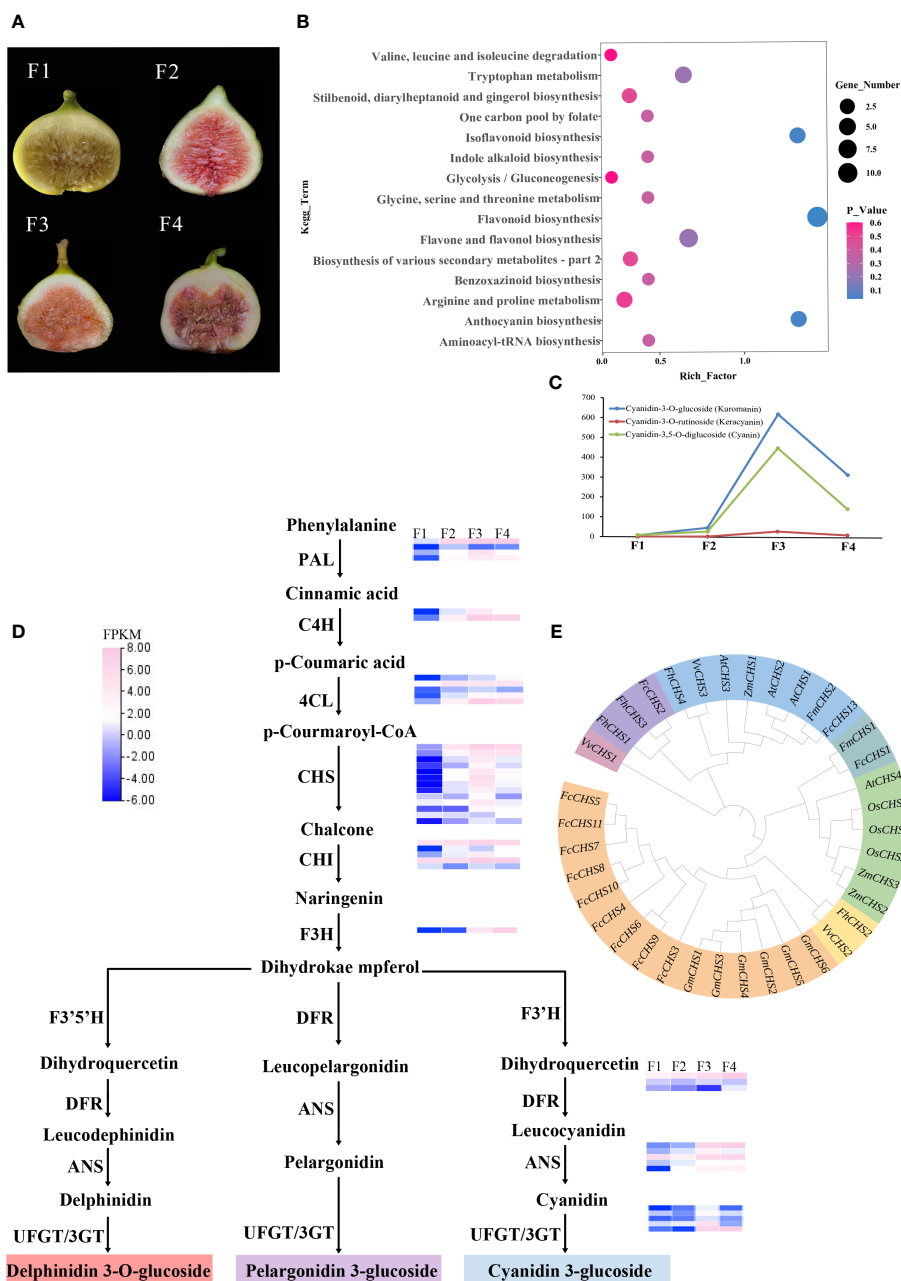


FIGURE 4

Transcriptome metabolism analysis of four different colors of *F. carica*. (A) The varieties of *F. carica* with different colored fruits. F1: "Orphan," F2: "Balaonai," F3: "Violette Solise," F4: "Bpjihon." (B) Pathway enrichment analysis of the differentially-accumulated metabolites in F1 vs. F3. (C) Changes in the levels of anthocyanins in the four different varieties of *F. carica*. (D) Flavonoids biosynthesis pathway and gene expression in the four varieties of *F. carica* with different colored fruits. Gene expression levels (log2-based FPKM) in different varieties are represented by color grading. (E) Phylogenetic tree of the CHS gene family in various species.

‘monooxygenase activity,’ and ‘tetrapyrrole binding’ were significantly enriched (Supplementary Figure 10). For F1 and F4, ‘cell death,’ ‘symbiotic-induced host-programmed cell death,’ ‘response to auxin,’ ‘secondary metabolic process,’ ‘secondary metabolite biosynthesis process,’ ‘chloroplast thylakoids,’ ‘neutral components of plasma membrane,’ ‘parts of plasma membrane,’ ‘thylakoids,’ ‘ADP binding,’ ‘heme binding,’ ‘oxygenase reductase activity,’ and ‘tetrapyrrole binding’ were significantly enriched (Supplementary Figure 11).

Differential gene expression analysis was also performed using KEGG enrichment. Significant enrichment was found in ‘plant-pathogen interaction,’ ‘plant hormone signal transduction,’ ‘phenylpropanoid biosynthesis,’ and ‘secondary metabolite biosynthesis’ between F1 and F2 (Supplementary Figure 12), ‘plant-pathogen interaction,’ ‘phenylpropane biosynthesis,’ and ‘biosynthesis of secondary metabolites’ between F1 and F3 (Supplementary Figure 13), and ‘plant-pathogen interaction,’ ‘phenylpropanoid biosynthesis,’ ‘flavonoid biosynthesis,’ and ‘biosynthesis of secondary metabolites’ between F1 and F4 (Supplementary Figure 14).

TABLE 1 Statistics for assembly and annotation of the *F. carica* genome.

Sequencing			
Sequencing platform	NovaSeq6000	Nanopore	Hi-C
Cleaned data (Gb)	26	42	62
Genome sequencing depth (×)	75	100	140
Assembly	Orphan	Dottato (Usai et al., 2020)	Horaishi (Mori et al., 2017)
Assembled genome size (Mb)	373,718,651	333,400,567	247,090,738
Sequence assigned to chromosomes (Mb)	366,336,389	266,522,563	–
Number of chromosomes	13	13	–
Max Contig (bp)	22,062,110	5,010,936	1,764,766
Min Contig (bp)	37,148	20,012	479
Contig N50 (bp)	9,781,938	823,517	166,092
BUSCO completeness (%)	96.2	93.3	90.5
Annotation			
GC content (%)	34.2		
Number of protein-coding genes	29,783		
Average gene length (bp)	3,111		
Average exon number per gene	5.2		

Differences in the secondary metabolites of *F. carica* varieties

Sequencing the transcriptome of the four varieties of *F. carica* suggested remarkable differences in the secondary metabolic processes of the species. The four differently colored fruits of *F. carica* were collected to understand the underlying mechanisms. The metabolic analysis conducted in this study identified 348 secondary metabolites, which included terpenes, flavonoids, and phenolic acids. Principal component analysis was performed on the data obtained from the gas chromatography-mass spectrometry to compare the differences in the metabolites of the ripe fruits of the four varieties of *F. carica*. Principal component analysis could easily distinguish between F1, F2, F3, and F4. Using the scoring graphs of PC1 and PC2, the compositions of the metabolites in the four samples could be distinguished. PC1 and PC2 were separated among the four groups of samples. PC1 and PC2 explained 35.8% and 25.08% of the total variance, respectively (Supplementary Figure 7A).

Differences in the levels of metabolites that affect fruit color in the four varieties of *F. carica*

Color is one of the critical characteristics considered for research related to improving horticultural plants. Fruit color is mainly affected by anthocyanins or carotenoids. The levels and types of metabolites may play a crucial role in determining the color of fruits in *F. carica*. The four fruit samples showed variations in the contents of terpenes, phenolic acids, and flavonoids. A total of 128, 129, and 111 differentially-accumulated secondary metabolites were identified between F1 and F2, F1 and F4, and F1 and F3, respectively. These included 51 metabolites that increased, and 76 that decreased between F1 and F2; 62 increased, and 67 decreased between F1 and F3; and 67 increased, and 44 decreased between F1 and F4. KEGG database was used to annotate the differentially-accumulated metabolites, and the results obtained indicated that the pathways that were mainly enriched included: ‘flavonoid biosynthesis,’ ‘anthocyanin biosynthesis,’ ‘isoflavone biosynthesis,’ ‘phenylalanine biosynthesis,’ and ‘isoflavone biosynthesis’ (Figure 4B, Supplementary Figure 7D).

Flavonoid biosynthesis pathway in the mature fruits of *F. carica*

Anthocyanins are flavonoids, water-soluble pigments that occur widely in plants and confer them with red, blue, and purple colors. A total of 94 flavonoids were detected through metabolomic analysis, of which anthocyanins were identified to be most closely associated with color. Three anthocyanins that were detected at significantly different levels in the assay were: cyanidin-3-O-glucoside (kuromanin), cyanidin-3-O-rutinoside (keracyanin), and cyanidin-3,5-O-diglucoside (cyanin), which were all up-regulated compared with those in F1. Amongst these, cyanidin-3-

O-glucoside was 600-fold higher in F3 than in F1, while cyanidin-3,5-O-diglucoside was 400-fold higher (Figure 4C).

Anthocyanins are mainly synthesized through the anthocyanin pathway, providing abundant natural pigments for different tissues and organs of plants. They are generally synthesized through the phenylalanine pathway and play a role in producing various derivatives through different metabolic pathways. In anthocyanin pathway, ten crucial gene families were identified, most located on chromosome 10, chromosome 12, and chromosome 13. The results of the analysis demonstrated that in the four varied colored *F. carica* fruits, the expression levels of *4CL*, *CHS*, *CHI*, *F3H*, *ANS*, and *UFGT* increased and were compatible with the rising anthocyanin content. In particular, there are notable variations in the expression levels of *CHS* and *UFGT* among different varieties of *F. carica*. Subsequently, we discovered 31 family members through *CHS* gene family analysis. Most of these genes were linked and localized on chromosome 10 of *F. carica*. Additionally, tandem repeat sequences were found, which aid in the *CHS* gene family's proliferation in *F. carica* (Figure 4D). A phylogenetic tree was also constructed based on the CDS sequences of *CHS* to determine the evolution of the *CHS* family in *F. carica*. It was discovered that *FcCHS3*, *FcCHS4*, *FcCHS5*, *FcCHS6*, *FcCHS7*, *FcCHS8*, *FcCHS9*, *FcCHS10* and *FcCHS11* are members of a subfamily. Notably, the *CHS* genes of *figs* show a distinct evolutionary tendency in comparison to other *Ficus* species (Figure 4E).

Discussion

F. carica is one of the first species to be domesticated, have significant economic and utilitarian importance, and is widely cultivated throughout Southwest Asia and the Middle East. Nonetheless, *F. carica* genetic research has been impeded by the absence of greater genome availability. We present a high-quality genome of "Orphan" with a contig N50 of 9.78 Mb and 366.34 Mb (98.02%) allocated to 13 chromosomes, which is valuable for understanding the genetics and evolutionary relationship, providing genomic resources and new insights into the breeding of *F. carica*. The integrity of the genome of "Orphan" as a reference was higher than that of *F. microcarpa* (contig N50 of 908kb), *F. hispida* (contig N50 of 492kb) (Zhang X. et al., 2020), and the previously reported genome of *F. carica*, 248 Mb size (contig N50 of 4.5 Kb) (Mori et al., 2017). In this study, nanopore sequencing (Belser et al., 2018) and high-throughput chromosome conformation capture (Jiao et al., 2017) were used for the assembly of genomes with a high quality. The complete sequence of the *F. carica* genome serves as a significant resource for future studies regarding the evolution and molecular breeding in *Ficus*.

In the first type of chromosomal number change mechanism, whole gene replication events are a common and significant chromosomal event that are necessary for the formation of new species or distinct phenotypes during evolution (Otto, 2007). The estimated divergence time between *Ficus* and *Morus* was ~120 MYA, and the differentiation time of *F. carica* and banyan *F. hispida* was ~34 MYA. A WGD event occurred roughly 2–3 MYA after *F. carica* and *Ficus* separated, according to Ks analysis

of the *F. carica* genome. Replication time gives redundant alleles unique or specialized activities, which may lead to the development of new regulatory mechanisms through genomic rearrangements. Chromosome fusion and breakage play equally important roles in the evolution of species. Chromosome number and ploidy will increase with biological evolution through whole genome duplication, polyploidization, and other processes. In addition, the genome may undergo diploidization to produce a small number of diploids, which will contribute to a decrease in chromosome number and ploidy. Related mechanisms include chromosome fusion and chromosome breakage (Mandakova et al., 2010; Soltis et al., 2016). Throughout their evolutionary history, *F. carica* have experienced chromosome fusion and fission with both *F. microcarpa* and *F. hispida*, including chromosome fusion and fission of chromosome 3 of *F. carica* with chromosomes 4 and 11 of *F. microcarpa*, and likewise in *F. carica* chromosome 3, which has also undergone chromosome fusion and fission with chromosomes 2 and 14 of *F. hispida*. It can be inferred from this that the chromosome 3 of *F. carica* is crucial for separating it from *Ficus* and creating a distinct species altogether.

An accidental event in a certain environment typically causes an increase or decrease in the total number of chromosomes in an organism during the course of its long-term evolution. This kind of chromosomal number alteration is referred to as the second kind. Progeny inherit a steady transmission of this change. Chromosomal recombination is the most interesting of these chromosomal events. Chromosome insertion, deletion, replication, inversion, translocation, and transposition are among the several modifications that can occur during chromosomal recombination. A chromosome can move from one chromosome to another, a process known as chromosomal translocation. Repeated regions on distinct chromosomes can also be attributed to chromosomal translocation. Chromosomes that experience translocation and inversion are likely to experience secondary recombination, and their structural changes will also impact the karyotype of the organism. All of these modifications have the potential to alter chromosomal numbers, which will promote the diversification and development of new species (Schubert and Lysak, 2011; Romanenko et al., 2019). Our research revealed that *F. carica* have experienced multiple chromosomal rearrangements across their evolutionary history, such as the inversion of chromosome 2 and the duplication of chromosomes 8 and 10. It's also important to note that the chromosome 9 of *F. carica* underwent both translocation and inversion. The 'chalcone metabolic process' and 'chalcone biosynthetic process' are the primary areas of enrichment for the functions of genes that undergo chromosomal recombination in *F. carica*. Chalcone Synthase (CHS) is the first enzyme in the pathway leading to the synthesis of plant flavonoids, which are not only extremely associated with plant fertility but also significantly impact plant resistance to pathogen infestation. The primary location of flavonoid synthesis in pollen is the chorioallantoic layer. From there, the flavonoids are transferred to the cyst cavity and ultimately to the pollen grain's outer wall, where they play a significant role. Thus, flavonoids are crucial for the development of pollen grains. Research has shown that the

examination of the flavonoid content in anthers will help to verify that the development of male sterility in *CHS-A* transgenic plants may be caused by the transcription of *CHS* in anthers. Male sterility was also observed in the transgenic plants, which were successfully genetically modified to modify the color of the flowers when the positive *CHsA* gene was introduced into *Petunia* (Shao and Xiao, 1996). While *F. carica* are dioecious plants, *F. microcarpa* is monoecious. The sex-specific characteristics of *F. carica* are most likely the result of chromosomal recombination events that occurred after *F. carica* separated from *F. microcarpa*.

Furthermore, the chromosomal reorganization gene functions also associated with ‘gametophyte development’, ‘sucrose synthase activity’, ‘sucrose biosynthetic process’, and ‘sucrose metabolic process’. It is worth mentioning that *F. carica* fruit trees produce edible fruits and are shorter in height than the other two *Ficus* species. The selection of *F. carica* for these genes associated with development and growth as well as sugar synthesis is consistent with the goal of their domestication. The examination of selection forces on *F. carica* and *F. hispida* provides additional evidence for the argument. *F. carica* have evolved primarily in response to ‘negative regulation of growth’, ‘negative regulation of developmental growth’, ‘reaction to cytokinin stimulus’, and ‘regulation of cellular biosynthetic process’ genes with associated functions being positively selected. It is evident that the typical fruit characteristics of *F. carica* compared to other *Ficus* species can be attributed to the WGD event and Chromosomal recombinations. The genomic information of *F. carica* may facilitate the analysis of the evolutionary process undergone by *Ficus* and help improve the understanding of the physiological and morphological diversity of these plants.

FhAG2, a region exclusive to males in the *F. hispida* genome, is only found in the male genome and is absent in the female genome before and during maturation and the inflorescence of the female flower. But in the case of female *F. carica* species, we compared similar genes. This gene with a specific deletion, and there is a chromosomal leap between this gene in *F. carica* and *F. hispida*. A phylogenetic tree was constructed by aligning the CDS and protein sequences of this gene, the results of which were inconsistent. The CDS alignment suggested that *F. carica* evolved relatively independently, whereas according to the protein alignment, *F. carica* and *F. hispida* were more closely related. The Ka/Ks values demonstrated that the differences may be caused by the selection pressure and this protein-coding gene. The observed behavior could potentially be explained by convergent evolution within species, as *F. carica* and *F. hispida* may have developed comparable structural features to adapt to similar ecological niches. Therefore, the proteins that *Ficus* and *F. carica* share are essential for regulating *F. carica* parthenogenesis. The expression levels of this gene were increased during the development of fertilized ovules because it was not expressed in the male *F. hispida* inflorescences. As a result, the gene shared by female plants of *F. carica* may be able to both stimulate the maturation of female flowers without pollination and replace the gene’s increased expression levels, which are necessary for the pollination process in *F. carica*. The edible portions of ripe *F. carica* are the receptacles. Hence, increased expression of this gene

may cause parthenocarpy in *F. carica*, and more research is required to determine the precise roles played by this gene.

F. carica is a species that is edible and useful in medicine since it contains a variety of bioactive chemicals. Nevertheless, the metabolic and biosynthetic pathways of these chemicals have been the subject of very few studies. The genomic, transcriptomic, and metabolomic data provided new insights into the biosynthetic processes in *F. carica*, with transcriptomic analysis revealing marked differences in the ‘signal transduction of plant hormones’ and ‘biosynthesis of secondary metabolites’. Secondary metabolite analysis showed that the ‘biosynthesis of anthocyanins’ demonstrated remarkable variations amongst the different varieties of *F. carica*. Due to the varying accumulation of anthocyanins, the four varieties of *F. carica* that were chosen for this investigation had remarkably diverse fruit colors. The ‘biosynthesis of anthocyanins’ was completed based on the flavonoid metabolic pathway, which is divided into two stages. The first stage involves *CHS*, *CHI*, *F3H*, and *F3’5’H*, it is the common pathway of flavonoid biosynthesis, and is called the pre-synthesis reaction of anthocyanin biosynthesis. The second stage involves the enzymes *DFR*, *ANS*, and *UFGT*, which are unique to anthocyanin biosynthesis, and this stage is called the late synthesis reaction of anthocyanin biosynthesis (Williams and Grayer, 2004; Petroni and Tonelli, 2011). Ten significant gene families, *PAL*, *C4H*, *4CL*, *CHS*, *CHI*, *F3H*, *F3’H*, *DFR*, *ANS*, and *UFGT* were found to be associated with the production pathway of *fig* anthocyanins in this study. The differential expression levels of genes related to anthocyanin biosynthesis were consistent with the contents of anthocyanins in *F. carica*.

In particular, the expression levels of *CHS* and *UFGT* demonstrated significant variations amongst the different varieties of *F. carica*, which was similar to potatoes (Cho et al., 2016) and jujubes (Zhang Q. et al., 2020), indicating that these genes play an essential role in the biosynthesis of anthocyanins (Wang et al., 2017). *F. carica* possessed 31 members of the *CHS* gene family, of which 13 were expressed at different levels. Additionally, we found a tandem repeat sequence that supports the *figs’ CHS* gene family amplification. Furthermore, different *fig* types have distinct *CHS* expression patterns. A phylogenetic tree constructed using the *CHS* proteins in *F. carica* and other species, showed that *FcCHS3*, *FcCHS4*, *FcCHS5*, *FcCHS6*, *FcCHS7*, *FcCHS8*, *FcCHS9*, *FcCHS10*, and *FcCHS11*, belonged to a single subfamily and the *CHS* of *F. carica* and other *Ficus* species has evolved along different trends.

The “anthocyanin biosynthesis pathway” was created using two different types of genes: regulatory genes, which control the expression patterns and levels of structural genes, and the structural genes, which encode the various essential enzymes involved in anthocyanin biosynthesis. The latter encoded transcription factors, primarily *WD40*, *bHLH*, and *MYB* (Schwinn et al., 2014). These genes are transcription factors that regulate *F. carica* biosynthesis, which is one of the main mechanisms underlying the diversity of *F. carica* fruits. One of the features of *Ficus* plants is the hidden head inflorescence. Studying the fruit diversity of this variety of *Ficus* plant, the linked genes that contribute to its diversity, and its selective preservation are crucial for future research because of its higher economic value.

In summary, the transcriptome and secondary metabolome analyses, along with the high-quality reference genome, offer valuable insights into the genome evolution and diversification of *figs*. Additionally, the data from this study offers important resources for genetic research as well as for *fig* and other *fig* plant improvement.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repository and accession number(s) can be found in the article/Supplementary Material. Theraw data generated through RNA sequencing have been deposited in the National Genomics Data Center here: "<https://ngdc.cncb.ac.cn/gsub/submit/bioproject/PRJCA016877>". The assembled genome was also uploaded to the National Genomics Data Center here: "<https://ngdc.cncb.ac.cn/gsub/submit/bioproject/PRJCA016848>".

Author contributions

YB: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – original draft. MH: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – original draft. CZ: Formal Analysis, Investigation, Methodology, Writing – original draft. SJ: Formal Analysis, Investigation, Methodology, Writing – original draft. LZ: Formal Analysis, Investigation, Methodology, Writing – original draft. ZY: Conceptualization, Investigation, Methodology, Writing – original draft. QS: Methodology, Resources, Writing – original draft. ZX: Methodology, Project administration, Supervision, Writing – review & editing. MZ: Methodology, Project administration, Supervision, Writing – review & editing.

References

- Ayuso, M., Carpena, M., Taofiq, O., Albuquerque, T. G., Simal-Gandara, J., Oliveira, M., et al. (2022). Fig "*Ficus carica* L." and its by-products: A decade evidence of their health-promoting benefits towards the development of novel food formulations. *Trends Food Sci. Technology*. 127, 1–13. doi: 10.1016/j.tifs.2022.06.010
- Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants*. 4, 879–887. doi: 10.1038/s41477-018-0289-4
- Castellarin, S. D., Matthews, M. A., and Gaspero, G. D. (2007). Water deficits accelerate ripening and induce changes in gene expression regulating flavonoid biosynthesis in grape berries. *Planta*. 227 (01), 101–112. doi: 10.1007/s00425-007-0598-8
- Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 34 (17), i884–i890. doi: 10.1093/bioinformatics/bty560
- Cho, K., Cho, K. S., Sohn, H. B., Ha, I. J., Hong, S. Y., Lee, H., et al. (2016). Network analysis of the metabolome and transcriptome reveals novel regulation of potato pigmentation. *J. Exp. Bot.* 67, 1519–1533. doi: 10.1093/jxb/erv549
- Czemmel, S., Heppel, S. C., and Bogs, J. (2012). R2R3 MYB transcription factors: key regulators of the flavonoid biosynthetic pathway in grapevine. *Protoplasma*. 249 (02), 109–118. doi: 10.1007/s00709-012-0380-z
- De, B. T., Cristianini, N., Demuth, J. P., and Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics*. 22 (10), 1269–1271. doi: 10.1093/bioinformatics/btl097
- Duenas, M., Perez-Alonso, J. J., and Santos-Buelga, C. (2008). Anthocyanin composition in fig (*Ficus carica* L.). *J. Food Compos. Anal.* 21 (02), 107–115.
- El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47 (D1), D427–D432. doi: 10.1093/nar/gky995
- Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20 (1), 238. doi: 10.1186/s13059-019-1832-y
- Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U S A*. 117 (17), 9451–9457. doi: 10.1073/pnas.1921046117
- Hamed, M., Khalifa, M. Y., El, Hassab, M. A., Abourehab, M. A., Al, kamaly, O., Alanazi, A. S., et al. (2023). The potential roles of *Ficus carica* extract in the management of COVID-19 viral infections: A computer-aided drug design study. *Curr. computer-aided Drug design*, 1875–6697. doi: 10.1155/2022/2044282
- Harzallah, A., Bhouri, A. M., Amri, Z., Soltana, H., and Hammami, M. (2016). Phytochemical content and antioxidant activity of different fruit parts juices of three *figs* (*Ficus carica* L.) varieties grown in Tunisia. *Ind. Crops Products*. 83, 255–267. doi: 10.1016/j.indcrop.2015.12.043
- Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics*. 36 (7), 2253–2255. doi: 10.1093/bioinformatics/btz891

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the Science and Technology Special Fund of Hainan Province (ZDYF2022XDNY149), the Domestic Cooperation Program of Shanghai Science and Technology Committee: Innovative Utilization of Global Tropical Fruit Germplasm Resources (22015810400), and the Developing Bioinformatics Platform in Hainan Yazhou Bay Seed Lab (B21HJ0001).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2023.1298417/full#supplementary-material>

- Ikegami, H., Habu, T., Mori, K., Hirata, C., Hirashima, K., and Tashiro, K. (2013). De novo sequencing and comparative analysis of expressed sequence tags from gynodioecious fig (*Ficus carica* L.) fruits: caprifig and common fig. *Tree Genet. Genomes*. 9, 1075–1088. doi: 10.1007/s11295-013-0622-z
- Istace, B., Friedrich, A., d'Agata, L., Faye, S., Payen, E., Beluche, O., et al. (2017). *de novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. *Gigascience*. 6 (2), 1–13. doi: 10.1093/gigascience/giw018
- Jiao, W. B., Accinelli, G. G., Hartwig, B., Kiefer, C., Baker, D., Severing, E., et al. (2017). Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res*. 27, 778–786. doi: 10.1101/gr.213652.116
- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 45 (D1), D353–D361. doi: 10.1093/nar/gkw1092
- Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptsomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 14 (4), R36. doi: 10.1186/gb-2013-14-4-r36
- Kislev, M. E., Hartmann, A., and Bar-Yosef, O. (2006). Early domesticated fig in the Jordan Valley. *Science*. 312, 1372–1374. doi: 10.1126/science.1125910
- Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* 34 (7), 1812–1819. doi: 10.1093/molbev/msx116
- Mandakova, T., Heenan, P. B., and Lysak, M. A. (2010). Island species radiation and karyotypic stasis in Pachycladon allopolyploids. *BMC Evolutionary Biol.* 10 (1), 367. doi: 10.1186/1471-2148-10-367
- Mawa, S., Husain, K., and Jantan, I. (2013). *Ficus carica* L. (Moraceae): Phytochemistry, traditional uses and biological activities. *Evidence-Based Complementary Altern. Med.* 2013, 974256. doi: 10.1155/2013/974256
- Mori, K., Shirasawa, K., Nogata, H., Hirata, C., Tashiro, K., Habu, T., et al. (2017). Identification of RAN1 orthologue associated with sex determination through whole genome sequencing analysis in fig (*Ficus carica* L.). *Sci. Rep.* 7, 41124. doi: 10.1038/srep41124
- Olga, D., Sanjit, S. B., Arina, D. O., Sarah, K. N., Marie, H., Neva, C. D., et al. (2017). *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. 356 (6333), 92–95. doi: 10.1126/science.aal3327
- Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell*. 131, 452–462. doi: 10.1016/j.cell.2007.10.022
- Petroni, K., and Tonelli, C. (2011). Recent advances on the regulation of anthocyanin synthesis in reproductive organs. *Plant Sci.* 181, 219–229. doi: 10.1016/j.plantsci.2011.05.009
- Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5 (3), e9490. doi: 10.1371/journal.pone.0009490
- Purnamasari, R., Winarni, D., Permanasari, A., Agustina, E., Hayaza, S., and Darmanto, W. (2019). Anticancer activity of methanol extract of *Ficus carica* leaves and fruits against proliferation, apoptosis, and necrosis in. *Cancer Informatics*. 18, 1176935119842576. doi: 10.1177/1176935119842576
- Romanenko, S. A., Lyapunova, E. A., Saidov, A. S., O'Brien, P., and Bakloushinskaya, I. (2019). Chromosome translocations as a driver of diversification in mole voles ellobius (Rodentia, mammalia). *Int. J. Mol. Sci.* 20 (18), 4466. doi: 10.3390/ijms20184466
- Ruprecht, C., Lohaus, R., Vanneste, K., Mutwil, M., Nikoloski, Z., Peer, Y., et al. (2017). Revisiting ancestral polyploidy in plants. *Sci. Advances*. 3 (7), 3793–3806. doi: 10.1126/sciadv.1603195
- Schubert, I., and Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* 27 (6), 207–223. doi: 10.1016/j.tig.2011.03.004
- Schwinn, K. E., Boase, M. R., Bradley, J. M., Lewis, D. H., Derolles, S. C., Martin, C. R., et al. (2014). MYB and bHLH transcription factor transgenes increase anthocyanin pigmentation in petunia and lisianthus plants, and the petunia phenotypes are strongly enhanced under field conditions. *Front. Plant Sci.* 5, 603. doi: 10.3389/fpls.2014.00603
- Seppely, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14
- Shao, L., and Xiao, S. H. (1996). The effect of chalcone synthase gene on flower color and fertility of transgenic plants. *Acta Botanica Sinica*. 038 (007), 517–524. doi: 10.1007/BF02951625
- Shirasawa, K., Yakushiji, H., Nishimura, R., Morita, T., Jikumaru, S., and Ikegami, H. (2020). The *Ficus erecta* genome aids Ceratocystis canker resistance breeding in common fig (*F. carica*). *Plant* 102 (6), 1313–1322. doi: 10.1111/tpj.14703
- Simsek, E., Kilic, D., and Caliskan, O. (2020). Phenotypic variation of fig genotypes (*Ficus carica* L.) in the eastern Mediterranean of Turkey. *Genetika*. 52 (3), 957–972. doi: 10.2298/GENSER2003957S
- Solomon, A., Golubowicz, S., Yablowicz, Z., Grossman, S., Bergman, M., Gottlieb, H. E., et al. (2006). Antioxidant activities and anthocyanin content of fresh fruits of common fig (*Ficus carica* L.). *J. Agric. Food Chem.* 54 (20), 7717–7723. doi: 10.1021/jf060497h
- Soltis, D. E., Visger, C. J., Marchant, D. B., and Soltis, P. S. (2016). Polyploidy: Pitfalls and paths to a paradigm. *Am. J. Botany*. 103 (7), 1146–1166. doi: 10.3732/ajb.1500501
- Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–9. doi: 10.1093/nar/gkl200
- Tanaka, Y., Sasaki, N., and Ohmiya, A. (2008). Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *Plant* 54(4), 733–749. doi: 10.1111/j.1365-313X.2008.03447.x
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7 (3), 562–578. doi: 10.1038/nprot.2012.016
- Treutter, D. (2010). Significance of flavonoids in plant resistance and enhancement of their biosynthesis. *Plant Biol.* 7 (06), 581–591. doi: 10.1055/s-2005-873009
- Usai, G., Mascagni, F., Giordani, T., Vangelisti, A., Bosi, E., Zuccolo, A., et al. (2020). Epigenetic patterns within the haplotype phased fig (*Ficus carica* L.) genome. *Plant*. 102 (3), 600–614. doi: 10.1111/tpj.14635
- Vangelisti, A., Zambrano, L. S., Caruso, G., Macheda, D., Bernardi, R., Usai, G., et al. (2019). How an ancient, salt-tolerant fruit crop, *Ficus carica* L., copes with salinity: a transcriptome analysis. *Sci. Rep.* 9, 2561. doi: 10.1038/s41598-019-39114-4
- Varet, H., Brillet-Gueguen, L., Coppee, J. Y., and Dillies, M. A. (2016). SARTools: A DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-seq data. *PLoS One* 11 (6), e0157022. doi: 10.1371/journal.pone.0157022
- Veberic, R., Colaric, M., and Stampar, F. (2008). Phenolic acids and flavonoids of fig fruit (*Ficus carica* L.) in the northern Mediterranean region. *Food Chem.* 106, 153–157. doi: 10.1016/j.foodchem.2007.05.061
- Vinson, J. A., Zubik, L., Bose, P., Samman, N., and Proch, J. (2005). Dried fruits: excellent in Vitro and in Vivo antioxidants. *J. Am. Coll. Nutr.* 24, 44–50. doi: 10.1080/07315724.2005.10719442
- Wang, Y., Tang, H., Debarr, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40 (7), e49. doi: 10.1093/nar/gkr1293
- Wang, S., Yang, C., Tu, H., Zhou, J., Liu, X., Cheng, Y., et al. (2017). Characterization and metabolic diversity of flavonoids in citrus species. *Sci. Rep.* 7, 10549. doi: 10.1038/s41598-017-10970-2
- Williams, C. A., and Grayer, R. J. (2004). Anthocyanins and other flavonoids. *Nat. Prod. Rep.* 21, 539–573. doi: 10.1039/b311404j
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24 (8), 1586–1591. doi: 10.1093/molbev/msm088
- Zhang, Y., Butelli, E., and Martin, C. (2014). Engineering anthocyanin biosynthesis in plants. *Curr. Opin. Plant Biol.* 19, 81–90. doi: 10.1016/j.pbi.2014.05.011
- Zhang, Z., Li, J., Zhao, X. Q., Wang, J., Wong, G. K. S., and Yu, J. (2006). KaKs_Calculator: calculating ka and ks through model selection and model averaging. *Genomics Proteomics Bioinf.* 4 (4), 259–263. doi: 10.1016/S1672-0229(07)60007-2
- Zhang, H., Wang, L., and Derolles, S. (2006). New insight into the structures and formation of anthocyanic vacuolar inclusions in flower petals. *BMC Plant Biol.* 6 (4), 29. doi: 10.1186/1471-2229-6-29
- Zhang, Q., Wang, L., Liu, Z., Zhao, Z., Zhao, J., Wang, Z., et al. (2020). Transcriptome and metabolome profiling unveil the mechanisms of Ziziphus jujuba Mill. peel coloration. *Food Chem.* 312, 125903. doi: 10.1016/j.foodchem.2019.125903
- Zhang, X., Wang, G., Zhang, S., Chen, S., Wang, Y., Wen, P., et al. (2020). Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell*. 183 (4), 875–889. doi: 10.1016/j.cell.2020.09.043



OPEN ACCESS

EDITED BY

Fei Shen,
Beijing Academy of Agricultural and Forestry
Sciences, China

REVIEWED BY

Zitong Li,
Commonwealth Scientific and Industrial
Research Organization (CSIRO), Australia
Aalt-Jan Van Dijk,
Wageningen University and Research,
Netherlands

*CORRESPONDENCE

Osva A. Montesinos-López

✉ osval78t@gmail.com

Jose Crossa

✉ j.crossa@cgiar.org

RECEIVED 18 October 2023

ACCEPTED 19 February 2024

PUBLISHED 04 March 2024

CITATION

Montesinos-López A, Crespo-Herrera L,
Dreisigacker S, Gerard G, Vitale P,
Saint Pierre C, Govindan V, Tarekegn ZT,
Flores MC, Pérez-Rodríguez P,
Ramos-Pulido S, Lillemo M, Li H,
Montesinos-López OA and Crossa J (2024)
Deep learning methods improve genomic
prediction of wheat breeding.
Front. Plant Sci. 15:1324090.
doi: 10.3389/fpls.2024.1324090

COPYRIGHT

© 2024 Montesinos-López, Crespo-Herrera,
Dreisigacker, Gerard, Vitale, Saint Pierre,
Govindan, Tarekegn, Flores, Pérez-Rodríguez,
Ramos-Pulido, Lillemo, Li, Montesinos-López
and Crossa. 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.

Deep learning methods improve genomic prediction of wheat breeding

Abelardo Montesinos-López¹, Leonardo Crespo-Herrera²,
Susanna Dreisigacker², Guillermo Gerard², Paolo Vitale²,
Carolina Saint Pierre², Velu Govindan²,
Zerihun Tadesse Tarekegn², Moisés Chavira Flores³,
Paulino Pérez-Rodríguez⁴, Sofía Ramos-Pulido¹,
Morten Lillemo⁵, Huihui Li⁶, Osva A. Montesinos-López^{7*}
and Jose Crossa^{2,4*}

¹Departamento de Matemáticas, Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI),
Universidad de Guadalajara, Guadalajara, Jalisco, Mexico, ²International Maize and Wheat
Improvement Center (CIMMYT), Texcoco, Estado de México, Mexico, ³Instituto de Investigaciones en
Matemáticas Aplicadas y Sistemas (IIMAS), Universidad Nacional Autónoma de México (UNAM),
Ciudad Universitaria, Ciudad de México, Mexico, ⁴Estudios del Desarrollo Rural, Economía, Estadística
y Cómputo Aplicado, Colegio de Postgraduados, Texcoco, Estado de México, Mexico, ⁵Department
of Plant Science, Norwegian University of Life Science (NMBU), Ås, Norway, ⁶State Key Laboratory of
Crop Gene Resources and Breeding, Institute of Crop Sciences and CIMMYT China Office, Chinese
Academy of Agricultural Sciences (CAAS), Beijing, China, ⁷Facultad de Telemática, Universidad de
Colima, Colima, Colima, Mexico

In the field of plant breeding, various machine learning models have been developed and studied to evaluate the genomic prediction (GP) accuracy of unseen phenotypes. Deep learning has shown promise. However, most studies on deep learning in plant breeding have been limited to small datasets, and only a few have explored its application in moderate-sized datasets. In this study, we aimed to address this limitation by utilizing a moderately large dataset. We examined the performance of a deep learning (DL) model and compared it with the widely used and powerful best linear unbiased prediction (GBLUP) model. The goal was to assess the GP accuracy in the context of a five-fold cross-validation strategy and when predicting complete environments using the DL model. The results revealed the DL model outperformed the GBLUP model in terms of GP accuracy for two out of the five included traits in the five-fold cross-validation strategy, with similar results in the other traits. This indicates the superiority of the DL model in predicting these specific traits. Furthermore, when predicting complete environments using the leave-one-environment-out (LOEO) approach, the DL model demonstrated competitive performance. It is worth noting that the DL model employed in this study extends a previously proposed multi-modal DL model, which had been primarily applied to image data but with small datasets. By utilizing a moderately large dataset, we were able to evaluate the performance and potential of the DL model in a context with more information and challenging scenario in plant breeding.

KEYWORDS

GBLUP model, genomic prediction, multi-modal deep learning model, machine learning methods, relationship matrices

Introduction

Wheat holds immense importance globally as a vital crop that serves as a staple food source for a significant portion of the world's population (Poland et al., 2012). It is cultivated in diverse agroclimatic regions and plays a critical role in ensuring global food security (FAO, 2021). The primary objective of wheat breeding programs is to develop superior varieties with enhanced traits such as higher yield potential, improved disease resistance, and better end-use quality. To expedite the breeding process and maximize genetic progress, genomics selection (GS) has emerged as a powerful tool (Crossa et al., 2017). In this context, genomic prediction has been extensively studied to enhance the efficiency of wheat breeding programs. It incorporates genomic relationship matrices to estimate the genetic variance and predict breeding values based on marker information.

Researchers have developed various statistical models to predict the performance of wheat lines based on genomic data. One fundamental and widely used model in genomic prediction is the Genomic Best Linear Unbiased Prediction (GBLUP) model, due in part to its simplicity and effectiveness in accounting for genetic relationships and accurately predict breeding values. GBLUP has demonstrated promising results in predicting complex traits in wheat, including yield, disease resistance, and quality attributes (Heffner et al., 2011; Poland et al., 2012; Rutkoski et al., 2016).

In recent years, deep learning models have gained attention for genomic prediction tasks in wheat. These models leverage the power of neural networks to learn complex patterns and relationships in genomic data (Crossa et al., 2017; Montesinos-López et al., 2018). The convolutional neuronal and the multilayer perceptron networks are the most common architecture applied in GS (Jiang and Li, 2020), and to reduce the number of weights to estimate during the training process more often a compressed version of the matrix of genomic relationship is used to feed the network instead of directly using the thousands of single nucleotide polymorphisms (SNP) available (Montesinos-López et al., 2018; Montesinos-López et al., 2021).

More recently, multi-modal deep learning models have emerged as an alternative that leverages multiple data modalities to improve prediction and analysis tasks (Liu et al., 2018). These models integrate multiple types of data inputs, such as genomic, phenotypic, and image environmental data, to improve prediction accuracy and robustness. By combining information from various sources, multi-modal models capture the interactions and correlations between different data modalities, leading to more accurate predictions and a better understanding of the underlying genetic architecture (Rahate et al., 2022).

Multi-modal deep learning has been explored and applied in diverse research fields, including the field of healthcare (Huang et al., 2020; Venugopalan et al., 2021; Kline et al., 2022; Stahlschmidt et al., 2022), agriculture (Danilevicz et al., 2021; Garillos-Manliguez and Chiang, 2021; Zhou et al., 2021), material sciences (Muroga et al., 2023), natural language processing (Morency and Baltrušaitis, 2017; Zadeh et al., 2018), social media

analysis (Balaji et al., 2021; Chandrasekaran et al., 2021), robotics and autonomous perception (Melotti et al., 2020; Duan et al., 2022).

For an early overview on deep multi-modal learning models see Ngiam et al. (2011) and Srivastava and Salakhutdinov (2012), and for a survey of recent advances in multi-modal machine learning see Ramachandram and Taylor (2017); Summaira et al. (2021) and Jabeen et al. (2023). In wheat genomic prediction, multi-modal deep learning models have been explored and applied as a promising approach (Kick et al., 2023; Montesinos-López et al., 2023). These studies have demonstrated the potential of multi-modal deep learning in enhancing the accuracy of genomic prediction for wheat traits.

Based on the previous considerations on how DL can be employed for genomic prediction in this study we follow a similar network structure as the previous study of Montesinos-López et al. (2023), up to the output layer. However, instead of directly combining the final outputs of individual networks from each modality to create the final output, we introduced an additional layer under a multi-layer perceptron network. This network has a similar architecture to the individual networks in each modality but with its own set of hyperparameters, which are also part of the tuning process. Furthermore, this study involves a moderately large dataset (4,464 wheat lines), allowing for a comprehensive evaluation of prediction accuracy. We compared the performance of our multi-modal deep learning model with the powerful GBLUP model, widely used in this field. This comparison enables us to assess the effectiveness of the multi-modal approach and its potential for enhancing genomic prediction accuracy in this specific context.

Materials and methods

Phenotypic data

The phenotypic data corresponds to the measurement of five traits (Yield, Germination, Heading, Height, and Maturity) in 4,464 wheat lines grown during the 2021/2022 crop season at the Norman E Borlaug Experiment Station, Ciudad Obregon (27°20' N, 109°54' W), Sonora, Mexico. The complete set of lines was tested under four different environments: (1) Beds with five irrigations (B5IR): genotypes were grown on raised beds with about 500 mm of available water and optimal sowing date during late November–early December, (2) Beds with two irrigations (B2IR): genotypes were grown on raised beds with about 250 mm of available water and optimal sowing date, (3) Bed Drought-Drip stress (BDRT): genotypes were grown on raised beds with about 120 mm of available water and optimal sowing date, and (4) Bed late heat stress (BLHT): genotypes were grown on raised beds with about 500 mm of available water and late sowing date (mid-February). Yield was measured in all environments, while Germination, Heading, Height and Maturity were determined in three out of four (B5IR, B2IR, and BDRT). Recently this data set was employed by Montesinos-López et al. (2023) for assessing the benefit of applying sparse phenotype field trials for genomic prediction at early testing generation of the population improvement (occurring at F_4 or F_5).

Genotypic data

The genotypic information comprised a total of 18,239 SNP markers. Genotyping was performed using the Genotyping-by-Sequencing (GBS) method, employing an Illumina HiSeq2500 sequencer at Kansas State University (Poland et al., 2012). Quality control was conducted using TASSEL v5.0 software (<https://tassel.bitbucket.io>). Raw data underwent filtration based on a minor allele frequency (MAF) cut-off of less than 5% and a missing data threshold of less than 50%. Subsequently, the HapMap file was converted into a numerical matrix to enable compatibility with the genomic prediction software. For the numerical representation, TASSEL assigned a value of 1 for homozygous major alleles, 0 for homozygous minor alleles, and 0.5 for heterozygous genotypes. To align the numerical matrix with the analysis tools utilized, substitution coding was applied, substituting the values with -1, 1, and 0, respectively. Finally, mean imputation was employed to address any missing values in the numerical matrix.

Statistical models

Bayesian GBLUP model

One of the statistical models used assumes that each response variable follows the relation:

$$Y_{ij} = \mu + E_i + g_j + gE_{ij} + \epsilon_{ij} \quad (1)$$

where Y_{ij} is the response variable for line j in environment i , μ is the general mean, E_i are the fixed effects of environment, g_j and gE_{ij} are the random effects of lines and random interaction effects of environment and line, respectively, and ϵ_{ij} are the random error terms assumed to be independent normal random variables with mean 0 and variance σ_e^2 . In addition, the random effects of lines and random genotype by environment interaction are assumed independently each other with the following distribution: $\mathbf{g} = (g_1, \dots, g_J)^T \sim N_J(\mathbf{0}_J, \sigma_g^2 \mathbf{G})$ and $\mathbf{gE} = (Eg_{11}, \dots, Eg_{IJ})^T \sim N_{IJ}(\mathbf{0}_{IJ}, \sigma_{Eg}^2 (\mathbf{I}_I \otimes \mathbf{G}))$ with $\mathbf{0}_J$ and \mathbf{I}_I the null vector of size J and the identity matrix of dimensions $I \times I$, and \otimes the Kronecker product.

A Bayesian estimation of these models was performed using a flat prior for the general mean and the fixed effects. For the variance components (σ_e^2 , σ_g^2 and σ_{Eg}^2) a scale inverse chi-squared distribution was employed. The model was implemented using the BGLR R package (Pérez-Rodríguez and de los Campos, 2014) with the default hyperparameter values.

DL model

The same information used in Equations 1 was employed to make predictions under the following multi-modal deep learning model (DL) with single output (Ouyang et al., 2014; Ramachandram and Taylor, 2017):

$$Y_{ij} = f(x_{ij}; \mathbf{W}) = f_O \left(w_0^{(O)} + x_{ij}^{*(L)T} \mathbf{w}_1^{(O)} \right) \quad (2)$$

where f_O is the output activation function with associated weights $w_0^{(O)}$ and $w_1^{(O)}$. $x_{ij}^{*(L)T}$ is the transpose of the vector with the neurons of last hidden layer ($x_{ij}^{*(L)}$) for a multilayer perceptron (MLP) neural network with L hidden layers, each layer with $N^{(l)}$ neurons and activation function f_l ($l = 1, \dots, L$), that use as input the concatenated outputs of the Q separately neural networks apply to each modality. That is, $x_{ij}^{*(L)}$ is computed recursively from:

$$\begin{aligned} x_{ij}^{*(l)} &= \left[x_{ij1}^{*(l)T}, \dots, x_{ijN^{(l)}}^{*(l)T} \right]^T = \left[f_l \left(z_{ij1}^{*(l)} \right), \dots, f_l \left(z_{ijN^{(l)}}^{*(l)} \right) \right]^T \\ &= \left[f_l \left(w_{01}^{(l)} + x_{ij}^{*(l-1)T} \mathbf{w}_{11}^{(l)} \right), \dots, f_l \left(w_{0N^{(l)}}^{(l)} + x_{ij}^{*(l-1)T} \mathbf{w}_{1N^{(l)}}^{(l)} \right) \right]^T \end{aligned}$$

where $\mathbf{W}_k^{(l)} = \left[w_{11}^{(l)}, \dots, w_{1N^{(l)}}^{(l)} \right]^T$ is the matrix of weights for layer l with $w_k^{(l)} = \left[w_{0k}^{(l)}, w_{1k}^{(l)} \right]^T$ for $k = 1, \dots, N^{(l)}$. Here $x_{ij}^{*(0)}$ is defined as $x_{ij}^{*(0)} = \left[x_{ij1}^{(L_1)T}, \dots, x_{ijQ}^{(L_Q)T} \right]^T$ with $x_{ijq}^{(L_q)}$ denoting the transpose of the vector $x_{ij(q)}^{(L_q)}$ that contain the outputs of the last hidden layer of the q -th MLP neural network (with L_q hidden layers, each layer with $N_q^{(l)}$ neurons and activation function $f_l^{(q)}$, $l = 1, \dots, L_q$) corresponding to the q -th modality ($q = 1, \dots, Q$), which in turn are computed recursively as:

$$x_{ij(q)}^{(l)} = \left[x_{ij1(q)}^{(l)}, \dots, x_{ijN_q^{(l)}(q)}^{(l)} \right]^T = \left[f_l^{(q)} \left(z_{ij1(q)}^{(l)} \right), \dots, f_l^{(q)} \left(z_{ijN_q^{(l)}(q)}^{(l)} \right) \right]^T$$

where $z_{ijk(q)}^{(l)} = w_{0k(q)}^{(l)} + x_{ij(q)}^{*(l-1)T} w_{1k(q)}^{(l)}$, $k = 1, \dots, N_q^{(l)}$, are linear transformations of the $N_q^{(l-1)}$ neurons in layer $l-1$ that define the neurons in layer l after applying the activation function $f_l^{(q)}$, $x_{ijk(q)}^{(l)} = f_l^{(q)}(z_{ijk(q)}^{(l)})$, $\mathbf{W}_{k(q)}^{(l)} = \left[w_{11(q)}^{(l)}, \dots, w_{1N_q^{(l)}(q)}^{(l)} \right]^T$ is the matrix of weights for the hidden layer l ($l = 1, \dots, L_q$) for the q -th neural network, $w_{k(q)}^{(l)} = \left[w_{0k(q)}^{(l)}, w_{1k(q)}^{(l)} \right]^T$ for $k = 1, \dots, N_q^{(l)}$, and $x_{ij(q)}^{(0)} = x_{ij(q)}$ are the inputs corresponding to q -th modality.

In the implemented models, all applied deep learning models are versions of Equation 2 that utilized a stacked residual network (ResNet) composed of 2 sequence layers (He et al., 2016). These were implemented with library TensorFlow in Python software, using a Batch_size value equal to 32, 48 epochs and the Adam optimizer (a stochastic gradient descend method to minimize the penalized loss function in DL) and using callback options of the fit keras function and specifying an adaptive exponential decay learning scheduler.

In all, for each modality (type of input) the number of units after the second hidden layer was equal to half of the units in the preceding layer, for example, for the neural network for the q -th modality,

$$N_q^{(l)} = \left\lfloor \frac{N_q^{(l-1)}}{2} \right\rfloor, \quad l = 2, \dots, L_q$$

where x denotes the largest integer less than x , and $N_q^{(1)}$ is the required number of units for the first hidden layer. Similarly, for the multilayer perceptron network after concatenating the outputs of the Q individual MLP neuronal networks, for a specified neurons in

its first layer ($N^{(1)}$), the neurons of the latter layers was taken as $N^{(l)} = \lfloor \frac{N^{(1)}}{2^{l-1}} \rfloor$, $l = 2, \dots, L$.

The rectified linear unit (ReLU) activation function was utilized in all hidden layers of the model, except for the output layer. For the output layer, a linear activation function was employed, assuming the conditional distribution of each trait follows a normal distribution. After each dense layer and prior to applying the activation function, a batch normalization layer was inserted. This layer help in approximately standardizing the outputs, ensuring a mean close to 0 and a standard deviation close to 1. For more detailed information, please refer to Figure 1.

For the training process, we employed an inner 10-fold cross-validation strategy. To expedite the training, only two out of the ten folds are utilized for validation. An early stopping rule is implemented through the callback option. The rule specifies monitoring the 'loss' function, with a mode of 'min' and a patience of 'Pat'. This rule checks whether the loss function on the training data stops decreasing at the end of each epoch. If it continues for an additional 'Pat' epochs, the training is halted.

To mitigate overfitting, dropout and L2 regularization were incorporated at each hidden layer, while only L2 regularization applied to the output layer. L2 regularization penalizes the loss function (e.g., sum of squared error loss) by adding the sum of squared weights multiplied by a regularization parameter (λ). This parameter controls the extent to which the weights are shrunk toward zero, reducing the model's complexity and preventing

excessive fitting to the training data. Dropout involves randomly setting a fraction of the weights to 0 at each training step.

Hyperparameters tuned in the experiment included learning decay (wd), patience values (Pat), dropout rate (DO), and regularization parameters (λ). The optimization of these hyperparameters was performed using the bayes_opt library with 50 iterations. The objective was to find the combination of hyperparameter values that minimized the mean squared error on the validation set. Table 1 provides a complete list of the hyperparameters and their corresponding search space.

The models were executed on a single computer node with 32 GB of RAM and 16 cores, together with a 20 GB GPU, and the experiments were conducted using Python version 3.8.10 and TensorFlow 2.11.0. On average, training each time a DL model with the specified characteristics described in the paper took approximately between 8 and 15 hours. In subsequent references within this manuscript, DL will be used to denote the specific deep learning (DL) model given in Equation 2, except in the LOEO evaluation where only the line effect is used.

Specifically, for the 5-fold cross-validation (5FCV) strategy described in the next section, the multi-modal DL Equation 2 was trained with 3 modalities corresponding to the information of the matrix design of environment (X_E), the genotype information ($X_L = Z_L L_G$) and the environment-genotype interaction information ($X_{EL} = Z_{EL} L_{EG}$), where Z_L and Z_{EL} are the matrix design of lines and the matrix design of the environment-line interaction, and L_G and L_{EG} are respectively the upper triangular

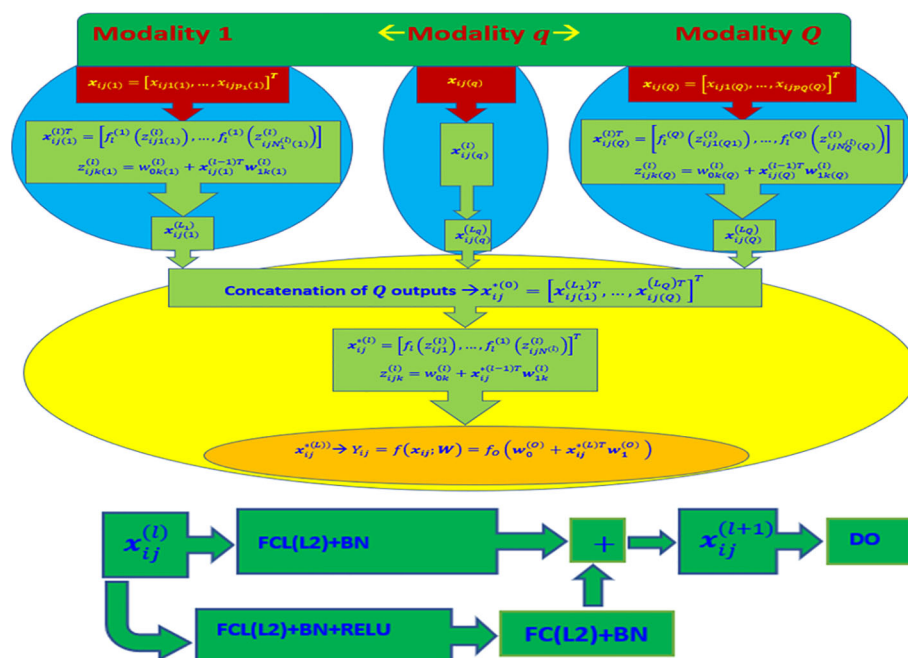


FIGURE 1

Top diagram: Multi-modal deep learning model (DL) with Q modalities (types of input). Bottom diagram: Stacked Residual Network (ResNet) composed of two sequential dense layers (FCL) applied in each MLP Neural Network. FCL(L2) + BN + ReLU denotes the successive application of a fully connected layer (FCL) with L2 regularization, batch normalization layer, and a ReLU activation function. Similarly, FCL(L2) + BN indicates the application of a fully connected layer with L2 regularization and batch normalization, while "DO" indicates the application of dropout regularization. The final output is produced by using the concatenated outputs of the Q networks as input to another MLP Neural Network. The output layer of this network consists of one neuron with a linear activation function and L2 regularization for its weights (concatenated outputs of all Q MLP Neural Networks + FCL + L2).

TABLE 1 Hyperparameters of the DL model and their respective domain space.

Hyperparameter	Notation	Bounds
Hidden layers for the MLP NN for each modality	L_1, L_2 and L_3	(1,4), (1,6) and (1,6)
Hidden layer for the MLP after concatenating the outputs of the NN of the 3 modalities	L	(0,4)
Number of neurons for the first layer in each modality	$N_1^{(1)}, N_2^{(1)}, N_3^{(1)}$	(0, 128), (1,1024) and (1, 1024)
Number of neurons for the first layer in the MLP after concatenating the outputs of the NN of the 3 modalities	$N^{(1)}$	(0, 200)
Regularization parameter for L2	λ	(1e-8,1e-2)
Dropout	DO	$(1 \times 10^{-4}, 0.5)$
Log weight decay	$lwd = \ln(wd)$	$(\ln(4 \times 10^{-5}), \ln(4 \times 10^{-1}))$
Patience	Pat	(0, 128)
Log learning rate	$lir = \ln(lr)$	$(\ln(1 \times 10^{-8}), \ln(1 \times 10^{-2}))$

part of the Cholesky decomposition of the genomic relationship matrix G ($G = L_G^T L_G$) and the upper triangular part of the Cholesky decomposition of the “environment-genomic” relationship matrix $G_{EG} = I_I \otimes G$ ($G_{EG} = L_{EG}^T L_{EG}$).

To evaluate the DL models for predicting the performance of an entire environment using the lines from all other environments (LOEO), the same DL model was employed. However, in this case, only the information of the matrix design of environment (X_E) and the genotype information were utilized as inputs. As a result, in the first predictor (GID) the DL is reduced to a single-modal DL model.

Assessment of prediction accuracy

Two strategies were used to evaluate and compare the models’ predictive performance. The first strategy, 5FCV, involved dividing the dataset into five balanced subsets. Four subsets were used for training the model, while the remaining subset was reserved for testing. This process was repeated, ensuring each subset served as the testing set once. The model’s performance was assessed by calculating the average Normalized Root Mean Squared Error (NRMSE) and Pearson’s correlation coefficient across all five partitions. The standard deviation was also computed to judge performance variability.

The second strategy, LOEO, is focused on predicting an entire environment using data from the other environments as training. During training, the models excluded the effects of environment (E) and the interaction between environment and lines (Eg). NRMSE and Pearson’s correlation coefficient were calculated for each predicted environment separately, allowing a detailed evaluation of the model’s performance in predicting specific environments.

By employing these strategies, the models’ predictive accuracy was assessed using NRMSE and Pearson’s correlation coefficient. The 5FCV approach provided an overall performance evaluation across the five cross-validation partitions, while LOEO enabled the evaluation of performance in individual environments.

Data availability

The phenotypic and genomic wheat data employed in this study can be downloaded from the following link <https://hdl.handle.net/11529/10548813> (Montesinos-López et al., 2023).

Results

The results are provided in three sections. First, for evaluating the prediction performance under tested lines in tested environments under a 5FCV, second, under tested lines in untested environments under the LOEO strategy and third, a summary of the hyperparameter values used in the trained models.

Tested lines in tested environments under a 5FCV strategy

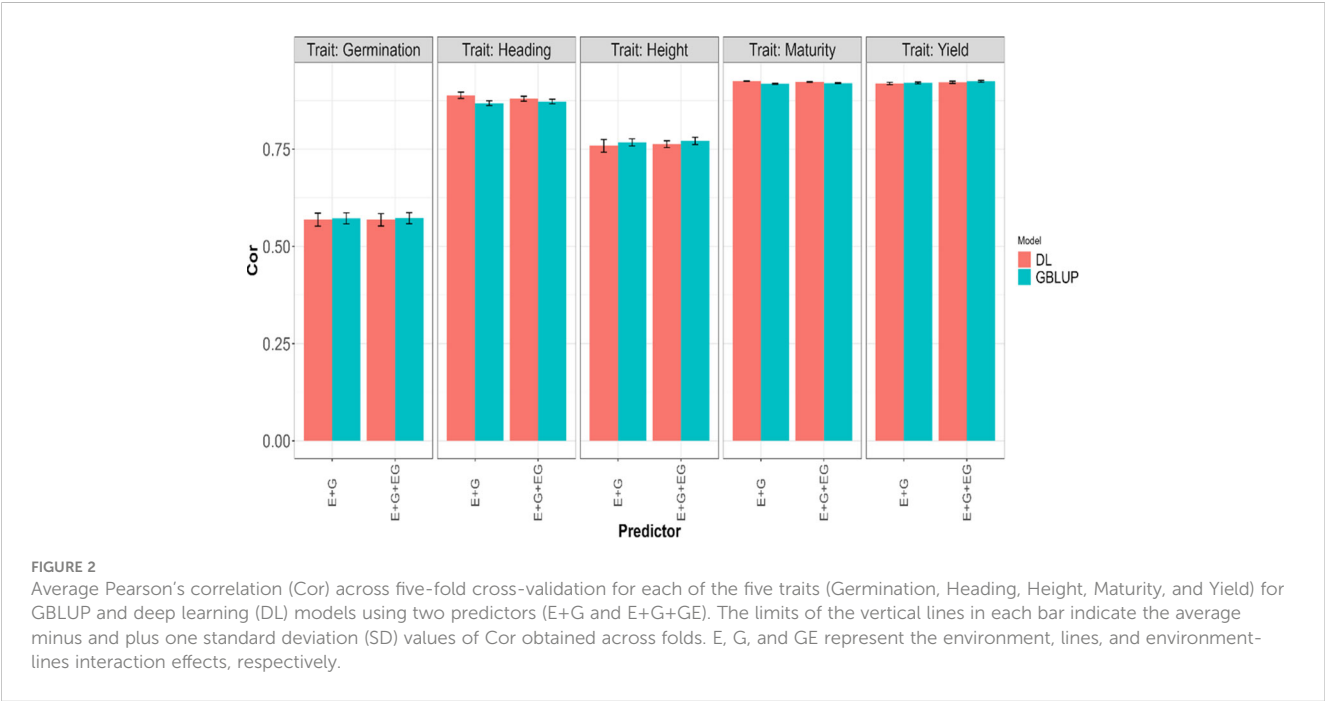
The fitted models for each of the four traits separately included the GBLUP Equation 1 and the deep learning Equation 2, along with sub-models of these primary models. Specifically, the first assessment of these models regarding its genomic prediction ability was conducted using the 5FCV strategy with the predictors $E + G + GE$ and $E + G$. The results are presented in Table 2 with the first, second and third columns indicating the model (GBLUP or DL), the trait and the predictor, respectively, and the last two columns the average and standard deviations values of the evaluated metrics (NRMSE and Cor). The results are also displayed in Figures 2 and 3. From Table 2, it can be observed that the GBLUP model performed best on average under the two evaluated metrics for three out of the five studied traits: Yield, Height, and Germination. The DL models showed an average NRMSE between 0.27% and 1.76% higher than the corresponding GBLUP models. However, the difference in performance was less pronounced for the Germination trait. In terms of the average correlation (Cor), the GBLUP model had values between 0.15% and 1.13% higher than those observed with the DL models. With this metric, the difference in performance was less pronounced for the Yield trait.

For Maturity and Heading, the DL models demonstrated better performance under the two evaluated metrics; the GBLUP model yielded an average NRMSE between 1.6% and 7.68% higher compared to the values obtained with the DL models, and in terms of the average Pearson’s correlation (Cor), the DL models provided between 0.33% and 2.33% higher values compared to those obtained with the GBLUP model. Furthermore, we can observe the GBLUP model exhibited a slightly better performance in all traits when using the predictor that involved environment, line, and environment-line interaction effects ($E+G+GE$) compared

TABLE 2 Average normalized root mean squared error of prediction (NRMSE) and average Pearson’s correlation (Cor) in a 5-fold cross-validation strategy when predicting each one of the five traits (Yield, Maturity, Height, Heading and Germination) with GBLUP and DL models using E+G and E+G+EG as predictors.

Model	Trait	Predictor	NRMSE (SD)	Cor (SD)
GBLUP	Yield	E+G	0.0932(0.0008)	0.9203(0.0024)
GBLUP	Yield	E+G+GE	0.0908(0.001)	0.9245(0.0023)
GBLUP	Maturity	E+G	0.0259(0.0002)	0.9181(0.001)
GBLUP	Maturity	E+G+GE	0.0256(0.0002)	0.9199(0.0008)
GBLUP	Height	E+G	0.0536(0.0006)	0.7674(0.0091)
GBLUP	Height	E+G+GE	0.0532(0.0006)	0.7711(0.0093)
GBLUP	Heading	E+G	0.0405(0.0006)	0.8683(0.006)
GBLUP	Heading	E+G+GE	0.0399(0.0006)	0.8725(0.0058)
GBLUP	Germination	E+G	0.082(0.0024)	0.5721(0.0141)
GBLUP	Germination	E+G+GE	0.082(0.0025)	0.5727(0.0142)
DL	Yield	E+G	0.094(0.0011)	0.9189(0.0032)
DL	Yield	E+G+GE	0.0923(0.0014)	0.922(0.0029)
DL	Maturity	E+G	0.0249(0.0002)	0.9249(0.0011)
DL	Maturity	E+G+GE	0.0252(0.0003)	0.9229(0.0012)
DL	Height	E+G	0.0545(0.0013)	0.7588(0.0163)
DL	Height	E+G+GE	0.0541(0.0006)	0.7628(0.0088)
DL	Heading	E+G	0.0376(0.0011)	0.8885(0.008)
DL	Heading	E+G+GE	0.0389(0.0008)	0.8798(0.0062)
DL	Germination	E+G	0.0823(0.0025)	0.5689(0.0166)
DL	Germination	E+G+GE	0.0825(0.0024)	0.5685(0.016)

SD represents the standard deviation of the metric across the folds.



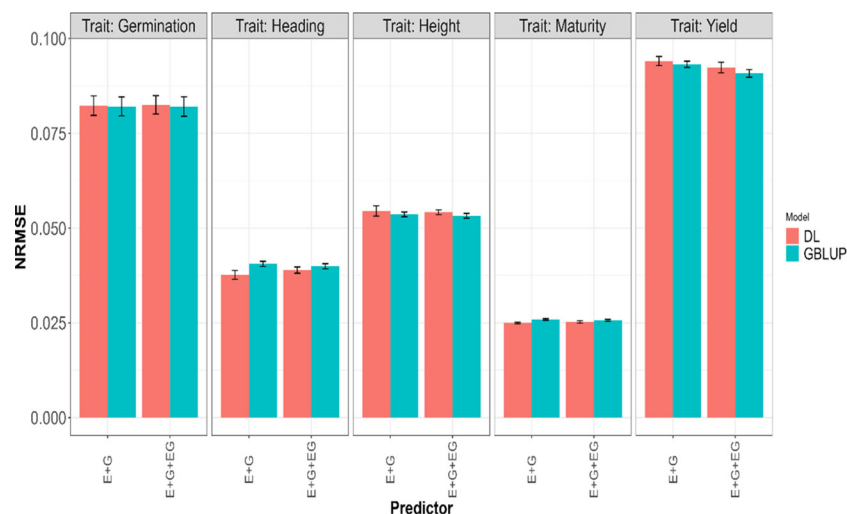


FIGURE 3

Average normalized mean squared error (NRMSE) across five-fold cross-validation for each of the five traits (Germination, Heading, Height, Maturity, and Yield) for GBLUP and deep learning (DL) models using two predictors (E+G and E+G+GE). The limits of the vertical lines in each bar indicate the average minus and plus one standard deviation (SD) values of Cor obtained across folds. E, G, and GE represent the environment, lines, and environment-lines interaction effects, respectively.

to the predictor with only the first two effects (E+G). However, with DL, this situation was observed only for the traits Yield, Height, and Germination with NRMSE, and only for the first two of these traits (Yield, Height) with the Cor metric. This indicates the importance of the environment-line interaction effect in the mentioned traits.

We observed an overlap of the intervals formed by subtracting and adding one standard deviation (SD) to the average metric values obtained in each model for each trait and predictor. From this, we can infer a very similar performance of both evaluated models in the 5FCV strategy. In fact, the average values across the five traits and all predictors (E+G, E+G+GE) for the average metrics presented in Table 2 are very similar, approximately 0.0587 for NRMSE and 0.81 for Cor.

Tested lines in untested environments LOEO strategy

In the LOEO strategy, the information of a complete environment was predicted with the rest of the environments in each trait. This was done with the GBLUP Equation 1 and DL Equation 2 under two predictors, the first with only line effect (G) and the second with environment plus line effect (E+G). The results are presented in Table 3 and Figures 4, 5. The first column indicates the trait to be predicted, the second column represents the predictor used, the third column denotes the environment to predict, and the last two columns display the NRMSE and Cor values obtained with the GBLUP and DL models, respectively.

Considering the 32 prediction scenarios, which correspond to all combinations of trait-predictor-environment (5 traits, 4 of these traits with three environments, and 1 trait with 4 environments, and 2 predictors (E and E+G)), we compared the performance of the models. In 11 out of 32 combinations, the DL model exhibited

smaller NRMSE values, while in another 11 out of 32 combinations, the DL model achieved higher Pearson's correlation values (Cor). Conversely, the GBLUP model outperformed the DL model in the remaining combinations.

Yield

GBLUP and DL showed better Cor performance when using only the line effect (G) compared to the predictor E+G. However, the NRMSE results exhibited a different pattern. In the GBLUP model, the G predictor outperformed E+G in three out of the four environments (B2IR, B5IR, and BLTH), while for the DL model, the more complex predictor (E+G) was only better than G predictor in one environment (B2IR) out of four. For this trait, the DL model outperformed the GBLUP model in two out of the four predicted environments. Specifically, the DL model performed better than the GBLUP model in the BLTH environment when considering the NRMSE metric, and in the B2IR environment when considering the Cor metric.

Maturity

GBLUP and DL showed better performance in terms of correlation (Cor) when using the E+G predictor compared to the G predictor. However, when considering the NRMSE metric, the results were opposite. The G predictor performed better in both models across all environments, exhibiting lower NRMSE values. Additionally, the DL model consistently showed higher correlation values than the GBLUP model in all environments. The DL model outperformed the GBLUP model in terms of NRMSE only in the B2IR environment.

TABLE 3 Normalized root mean squared error of prediction (NRMSE) and average Pearson’s correlation (Cor) in LOEO evaluation strategy when predicting each one of the five traits (Yield, Maturity, Height, Heading and Germination) with GBLUP and DL models.

Trait	Model		GBLUP		DL	
	Predictor	Env	NRMSE	Cor	NRMSE	Cor
Yield	G	B2IR	0.2244	0.1151	0.2344	0.0688
Yield	E+G	B2IR	0.58	0.196	0.2226	0.2072
Yield	G	B5IR	0.3678	0.2025	0.3679	0.1323
Yield	E+G	B5IR	0.3757	0.2242	0.3679	0.2183
Yield	G	BDRT	0.3343	0.1682	0.3345	0.0444
Yield	E+G	BDRT	0.1183	0.2004	0.3366	0.1612
Yield	G	BLHT	0.1087	0.1979	0.108	-0.0262
Yield	E+G	BLHT	0.137	0.3071	0.1094	0.259
Maturity	G	B2IR	0.065	0.4312	0.066	0.1832
Maturity	E+G	B2IR	0.1242	0.6294	0.0614	0.6697
Maturity	G	B5IR	0.114	0.5775	0.1133	0.6127
Maturity	E+G	B5IR	0.1092	0.5846	0.1146	0.6216
Maturity	G	BDRT	0.0789	0.3197	0.0801	0.2058
Maturity	E+G	BDRT	0.0307	0.5376	0.076	0.6061
Height	G	B2IR	0.0482	0.2779	0.0502	0.0721
Height	E+G	B2IR	0.0888	0.3433	0.0482	0.29
Height	G	B5IR	0.1178	0.2243	0.1171	0.1943
Height	E+G	B5IR	0.0873	0.2493	0.1189	0.2629
Height	G	BDRT	0.1392	0.1805	0.1392	0.0875
Height	E+G	BDRT	0.1047	0.2097	0.1403	0.1894
Heading	G	B2IR	0.0708	0.5594	0.0754	0.5039
Heading	E+G	B2IR	0.1232	0.7642	0.0568	0.8158
Heading	G	B5IR	0.1197	0.7412	0.121	0.7732
Heading	E+G	B5IR	0.1097	0.7558	0.1202	0.7894
Heading	G	BDRT	0.097	0.4528	0.1051	0.4359
Heading	E+G	BDRT	0.0512	0.6449	0.0953	0.6551
Germination	G	B2IR	0.1004	0.0907	0.1012	0.0447
Germination	E+G	B2IR	0.0723	0.0895	0.1002	0.0352
Germination	G	B5IR	0.0735	0.0308	0.0726	0.0184
Germination	E+G	B5IR	0.0651	0.0314	0.073	0.0134
Germination	G	BDRT	0.1727	0.0685	0.1728	0.0357
Germination	E+G	BDRT	0.1823	0.068	0.169	0.1063

The best predictor (G or G+E) for each combination (model/trait) is indicated in bold.

Height

The GBLUP model displayed better performance with the E+G predictor compared to the G predictor in two out of three environments for NRMSE and in all environments for Cor metric. However, the DL model exhibited a different pattern. For

NRMSE, the G predictor outperformed E+G in two out of the three environments, while for Cor, the DL model achieved better performance with the E+G predictor in all environments. When comparing the models, the DL model showed better NRMSE performance in the B2IR environment, while the GBLUP model outperformed in the other environments. In terms of correlation

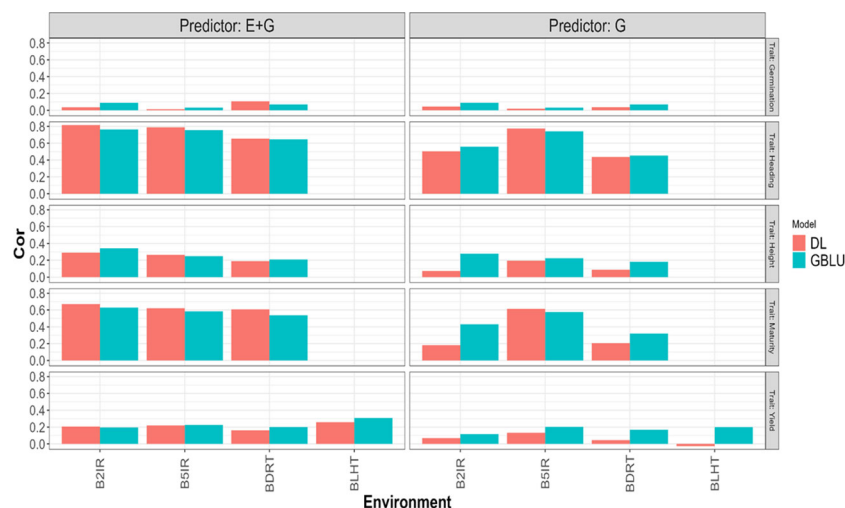


FIGURE 4

Pearson's correlation obtained in each environment when applying LOEO strategy for each of the five traits (Germination, Heading, Height, Maturity, and Yield) for GBLUP and multi-modal deep learning (DL) models using two predictors (G and E+G).

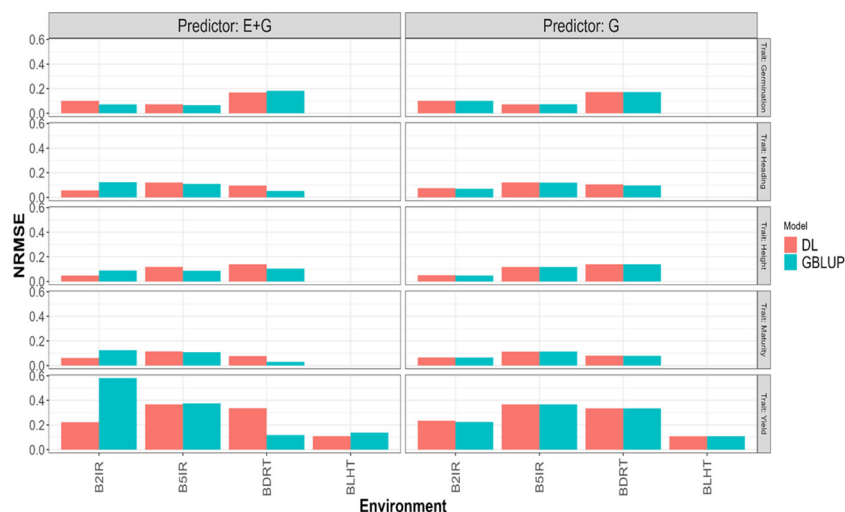


FIGURE 5

Normalized mean squared error (NRMSE) obtained in each environment when applying LOEO strategy for each of the five traits (Germination, Heading, Height, Maturity, and Yield) for GBLUP and multi-modal deep learning (DL) models using two predictors (G and E+G).

(Cor), the DL model exhibited better performance in the B5IR environment, while in the rest of environments the GBLUP model was superior.

Heading

The GBLUP model performed better with the E+G predictor compared to G in two out of the three environments for NRMSE and in all environments for Cor metrics. However, the DL model consistently showed better performance with the E+G predictor in terms of both NRMSE and Cor in all environments. In this case, the

DL model outperformed the GBLUP model in all environments when considering the Cor metric, and for the NRMSE metric, the DL model was better in only one environment (B2IR).

Germination

Both models showed better performance with the E+G predictor compared to the G predictor in two out of three environments in terms of NRMSE. However, the results were opposite in terms of Cor, where the G predictor exhibited better performance in the other two environments. In this case, the DL model outperformed the GBLUP

model in the BDRT environment for both NRMSE and Cor metrics, and in the remaining two environments the GBLUP was better.

Summary of the hyperparameter values used in the trained models

A summary of the optimal hyperparameter values used in the trained models for the 5FCV and LOEO evaluation strategies is provided in [Tables 4](#) and [5](#). The descriptions of [Table 4](#) are:

- For the modality corresponding to environment effects (E), the optimal number of hidden layers more frequently found across the 5 partitions by the Bayesian optimization was 1 and 2 for models with predictor E+G and E+G+GE, respectively. This pattern was observed in the Germination and Height. In Heading and Maturity, the most frequently observed optimal number of hidden layers were 2 for the E+G predictor and 3 for the E+G+GE predictor. For Yield, the optimal number of hidden layers varied, with 1 being the most frequently observed for the E+G predictor, and 3 being the most frequently observed for the E+G+GE predictor. Regarding the optimal number of units, for Germination and Height, the most frequently observed values were 128 units for the E+G predictor and 89 units for the E+G+GE predictor. For Yield, Maturity, and Heading with predictor E+G+GE the units required were 60, and were 128, 114 and 114 for the same traits but under predictor E+G.
- For the modality corresponding to the Line effect ($Z_L \times L_G$), the most frequently observed number of units was around 796 units for all traits in the model with the predictor E+G+GE. For the predictor E+G, the most frequently observed number of units varied across the traits, with 179,
- 183, 302, 302, and 472 units for the Yield, Height, Maturity, Heading, and Germination, respectively. Regarding the hidden layers in this modality, 3 and 1 were the most frequently observed values used in the models with both predictors (E+G and E+G+GE) for the Heading and Maturity traits. For the Height and Yield, regardless of the predictor (E+G and E+G+GE), the most frequently observed value was 1. Lastly, for Germination, the most frequently observed values for the number of hidden layers found by Bayesian optimization across the 5 partitions (5FCV) were 6 for the E+G predictor and 1 for the E+G+GE predictor.
- For the line-environment interaction modality effect, in all traits the most frequently optimal number of hidden layers observed was 1, and the corresponding optimal number of units varied depending on the trait. For Yield, Maturity, and Heading, the most frequently observed optimal number of units was 285, and for Germination and Height, the most frequently observed optimal number of units was 869.
- For 3 out of the 5 traits (Yield, Germination, and Height), in many of the folds, the DL model did not require hidden layers after the concatenation of the individual neural networks ($n_{HLB_2}=0$) when using the predictor E+G. In cases where more than one hidden layer was required, the most frequently observed optimal number of units (N_2) was 200 and 100 for Yield, and approximately 100 for Height and Germination. For the other two traits, the required number of hidden layers was 3. For the model using the predictor E+G+GE, the most frequently observed number of hidden layers was 2 for three traits (Yield, Maturity, and Heading), and 1 for Germination and Height. For model with predictor E+G+GE, the more often hidden layers observed were 2 for traits Yield, Maturity and Heading, and for these three traits the most

TABLE 4 Summary of the hyperparameter values used in the DL models for the 5-fold cross-validation (5FCV) performance evaluation strategy.

Trait	Predictor	l	llr	lwd	DO	$N_1^{(1)}$	$N_2^{(1)}$	$N_3^{(1)}$	$N^{(1)}$	L_1	L_2	L_3	L	Pat
Yield	E+G	0.0044	-4.8609	-0.9982	0.0046	1	1		0	128	179		200,100	1
Yield	E+G+GE	0.0003	-5.0740	-1.1720	0.1943	3	1	1	2	60	796	285	32	120
Maturity	E+G	0.0046	-7.4161	-5.6717	0.3997	2	3		3	114	302		76	42
Maturity	E+G+GE	0.0019	-5.5868	-1.5065	0.2521	3	1	1	2	60	796	285	32	120
Height	E+G	0.0078	-4.6052	-0.9163	0.0001	1	1		0	128	183		1	1
Height	E+G+GE	0.0045	-4.7462	-3.1049	0.3120	2	1	1	1	89	797	869	108	35
Heading	E+G	0.0023	-6.2648	-4.1796	0.1873	2	3		3	114	302		76	42
Heading	E+G+GE	0.0002	-5.1912	-1.2001	0.2250	3	1	1	2	60	796	285	32	120
Germination	E+G	0.0088	-4.6052	-0.9163	0.2001	1	6		0	128	472		200	128
Germination	E+G+GE	0.0075	-4.8166	-4.2082	0.3415	2	1	1	1	89	797	869	108	35

The first two columns indicate the trait and the predictor used in the evaluation. Columns 3 to 6 represent the average values of the regularization (l), the logarithm of the learning rate (llr), the logarithm of the weight decay (lwd) and the dropout rate (DO), respectively. In the columns 7 to 14 the most frequently observed optimal values (mode) across the 5 partitions in the 5-fold cross-validation (5FCV) for the hidden layers (L_1, L_2, L_3) and the number of units ($N_1^{(1)}, N_2^{(1)}, N_3^{(1)}$) in the respective networks for each modality in the model. These columns also include the information of the number of hidden layers (L) and number of units ($N^{(1)}$) for the network created by concatenating the outputs of the individual networks before the output layer. The final column indicates the most frequently observed optimal value for the patience (Pat) hyperparameter registered in the early stopping criteria across the partitions.

TABLE 5 Summary of the hyperparameter values used in the DL models for the LOEO performance evaluation strategy.

Trait	Predictor	l	llr	lwd	DO	L_1	L_2	L	$N_1^{(1)}$	$N_1^{(2)}$	$N^{(1)}$	Pat
Yield	GID	0.0046	-5.6924	-6.8825	0.3364		5			552		51
Yield	GID+Env	0.0099	-4.6052	-0.9163	0.0001	1	1	0	128	183	107	1
Maturity	GID	0.0039	-7.0800	-6.6678	0.4942		1			709		30
Maturity	GID+Env	0.0030	-6.4791	-4.0866	0.2665	2	3	3	114	302	76	42
Height	GID	0.0094	-4.7878	-6.3700	0.2604		1			101		1
Height	GID+Env	0.0100	-4.6052	-0.9163	0.0001	4	1	4	128	905	140	128
Heading	GID	0.0085	-6.6059	-8.1575	0.2615		2			101		110
Heading	GID+Env	0.0015	-5.5421	-5.5715	0.1333	4	1	0, 3, 4	128	302	NA, 76, 119	37
Germination	GID	0.0028	-5.0766	-3.2896	0.2616		6			127		1
Germination	GID+Env	0.0064	-6.4791	-4.0866	0.2665	2	3	3	114	302	76	42

The first two columns indicate the trait and the predictor used in the evaluation. Columns 3 to 6 represent the average values of the regularization (l), the logarithm of the learning rate (llr), the logarithm of the weight decay (lwd) and the dropout rate (DO), respectively. In the first 6 columns of the last 7 columns correspond to the most frequently observed optimal values (mode) across the predicted environments for the hidden layers (L_1, L_2) and the number of units ($N_1^{(1)}, N_2^{(1)}$) in the respective networks for each modality in the model. In these columns also is included the information of the number of hidden layers (L) and number of units ($N^{(1)}$) for the network created by concatenating the outputs of the individual networks before the output layer. The final column indicates the most frequently observed optimal value for the patience (Pat) hyperparameter registered in the early stopping criteria across the partitions.

frequently optimal number of units was 32. For Germination and Height, the most frequently number of hidden layers used was 1 and the most frequently optimal number of units was 797.

- The most frequently optimal values for the patience hyperparameter (Pat) ranged between 1 and 128 across the 5 traits and the two evaluated predictors. The most observed value was 120. Regarding the rest of the hyperparameters, the regularization parameter (l) ranged between 0.0003 and 0.0088 across all traits and predictors, with an average optimal value of 0.004. The logarithm of the learning rate (llr), logarithm of the weight decay (lwd), and dropout regularization (DO) values ranged between (-7.4161, -4.6052), (-5.6717, -0.9163), and (0.0001, 0.3997) respectively. The average values of the most frequently observed values were -5.3167 for llr , -2.3874 for lwd , and 0.2117 for DO .

When predicting a complete environment using the rest (LOEO), the most frequently optimal values of the integer hyperparameters (hidden layers, units, and patience) for the trained DL models are presented in Table 5. Additionally, the table includes the average values of the optimal real-valued hyperparameters (across environments) for the described Equation 2. While there are variations in the configurations of the NN models across traits and predictors, certain patterns can be observed. Across all traits and predictors, the average optimal values (across predicted environments) for the regularization parameter (l), the logarithm of the learning rate (llr), the logarithm of the weight decay (lwd), and dropout regularization (DO) fall within the intervals (0.0015, 0.01), (-7.08, -4.6051), (-8.1574, -0.9162), and (0.0001, 0.4942), respectively. The average values of these average optimal values are approximately in the middle of these intervals.

We observed the following patterns for the models with different predictors.

- For models with the predictor G, the most frequently optimal number of hidden layers (column L_2) for the corresponding neuronal networks were 5, 1, 1, 2, and 6 for traits Yield, Maturity, Height, Heading, and Germination, respectively. The corresponding number of units ($N_2^{(1)}$) were 552, 709, 101, 101, and 127, with none reaching the upper bound of 1024 set in the search bounds (Table 1).
- For models with predictor E+G, in the individual NN of the modality of GID effect, the most frequently optimal number of hidden layers was 1 for Yield, Height, Heading, and 3 for traits Germination and Maturity. The corresponding most frequently optimal number of units used across the predicted environments were 183, 905, 302, 302, and 302 for Yield, Height, Germination, Maturity, and Heading, respectively.
- For models with the predictor E+G, in DL model with the modality corresponding to the Env effect (E), the most frequently optimal number of hidden layers were 4 in two traits (Heading and Height), 2 in two traits (Germination and Maturity), and 1 in the remaining trait (Yield). The corresponding most frequently optimal values of units were 128, 128, 114, 114, and 128 for traits Heading, Height, Germination, Maturity, and Yield, respectively. In Yield, no hidden layers were used in most of the fitted models after concatenating the outputs of the NNs of the involved inputs (Env and G), and for Heading were required 0, 3 and 4 hidden layers for the three predicted traits with none (not apply), 76 and 119 units, respectively. However, for Maturity, Height, and Germination, the most frequently optimal values for the number of hidden layers were 3, 4,

and 3, respectively, as determined by the Bayesian optimization algorithm. When required at least one hidden layer ($n_{HLB} > 1$) in the trained model for predicting an environment, the most frequently optimal number of units for the first layer after concatenating the outputs of the individual NNs for Env and GID, were 107, 76, 140, 69, and 76 for Yield, Maturity, Height, Heading, and Germination, respectively.

An impact evaluation of the data size on accuracy

An evaluation of the impact of the dataset size in the accuracy prediction but with less computational time was done using a reduced search space bounds as the specified in the shared code example. The search space includes the interval [1,2] for all hidden layers, [4,8] for the units of the environment effect, [32, 128] for the units in the line and effect, and the same interval for the units in the hidden layer for the MLP after concatenating the outputs of the neural networks of the two modalities (Environments and Lines effects). Additionally, we utilized the same search space for the rest of the hyperparameters, as described in Table 1.

This evaluation for both models (DL and GBLUP both with predictor E+G) was conducted by retaining 5%, 10%, 50%, 66.6%, and 80% (Percentage_tr) of the dataset for the training set, with the remainder used for the testing set. In all cases, we adhered to the spirit of the K-fold cross-validation strategy. For the first two cases (20-Fold and 10-Fold), the training and testing roles were inverted (1 fold for training and the rest of the folds for testing). For the last three cases, the traditional K-fold cross-validation strategy (2-Fold,

3-Fold, and 5-Fold) was implemented, where K-1 subsets were used for training, and the remaining subset was used for testing. Furthermore, the K-Folds in the third and fourth cases were repeated two times to obtain more representative results.

The obtained results are summarized in Figures 6 and 7, where the height of the bars represents the average metric values across folds. The vertical lines within each bar indicate the average minus and plus one standard deviation (SD) values of Cor obtained across folds. In the first of these figures (Figure 6), a deterioration in the normalized root mean squared error is observed as the training size decreases (moving right to left on the Percentage_tr axis) in both explored models. This deterioration is more pronounced in the Heading and Maturity traits. However, in all traits, this effect tends to be slightly smaller in the GBLUP model. A similar behavior is observed in Figure 7 concerning the average Pearson's correlation. These results are also very similar to those reported in the 5FCV strategy with the larger explored search space.

Discussions

In this study, we utilized and expanded upon a recently proposed multi-modal DL model (Montesinos-López et al., 2023) for genomic prediction. Our extended model incorporated a neural network that takes as input the concatenated outputs of the individual NNs for each modality (E, G, and GE, for example). The improved performance of the DL models can be attributed, in part, to the novel architecture employed and to the availability of a moderately larger dataset.

Within the application of multi-modal deep learning in the context of genomic selection, it is important to take advantages of the virtues of multi-modal deep learning:

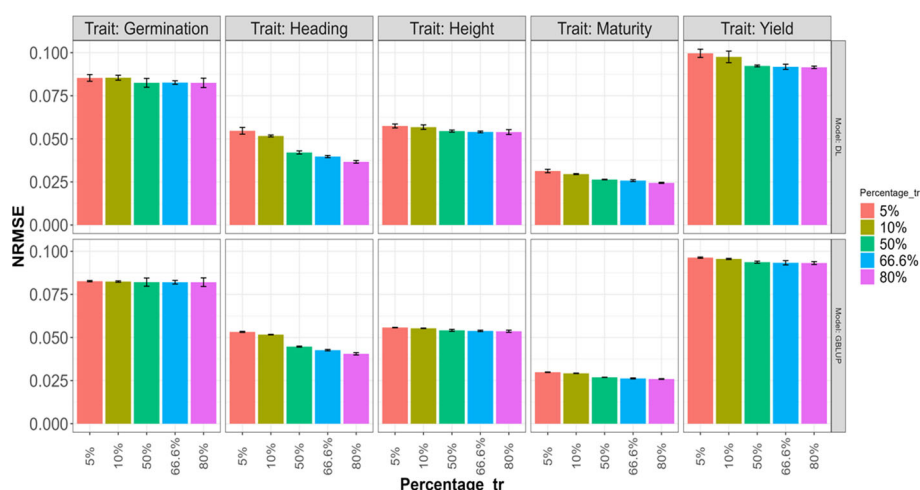


FIGURE 6

Average normalized mean squared error (NRMSE) across folds for each of the five traits (Germination, Heading, Height, Maturity, and Yield) for GBLUP and deep learning (DL) models using the predictor E+G. Percentages represent the portion of the dataset used for training. The bars for the first two values (5% and 10%) correspond to results in a 20-Fold and 10-Fold cross-validation strategy, with one-fold for training and the rest for testing. The remaining bars for the last three Percentage values correspond, respectively, to the traditional 2-Fold, 3-Fold, and 5-Fold cross-validation strategies, with the first two being repeated two times.

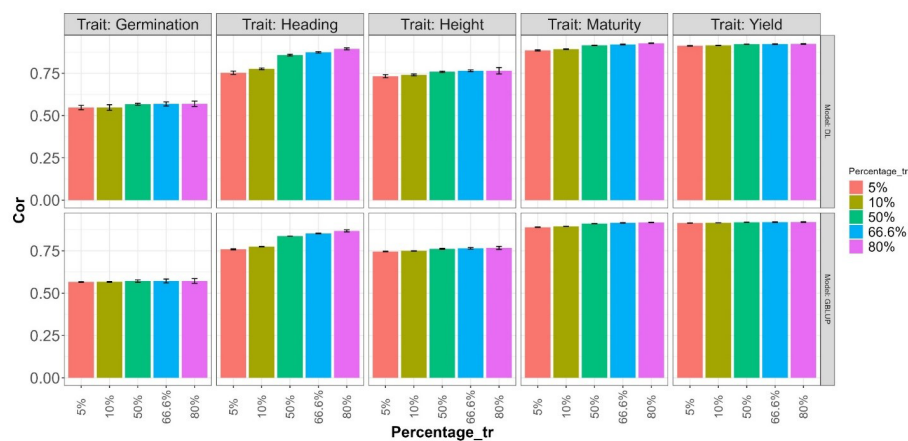


FIGURE 7

Average Pearson's correlation (Cor) across folds for each of the five traits (Germination, Heading, Height, Maturity, and Yield) for GBLUP and deep learning (DL) models using the predictor E+G. Percentages represent the portion of the dataset used for training. The bars for the first two values (5% and 10%) correspond to results in a 20-Fold and 10-Fold cross-validation strategy, with one fold for training and the rest for testing. The remaining bars for the last three Percentage values correspond, respectively, to the traditional 2-Fold, 3-Fold, and 5-Fold cross-validation strategies, with the first two being repeated two times.

- (1) Enhanced representation learning, by integrating different modalities, since multi-modal deep learning can learn richer and more comprehensive representations of data. This allows for a more holistic understanding of the input, capturing both complementary and redundant information across modalities.
- (2) Improved performance because multi-modal deep learning outperforms single-modal approaches in various tasks, including image captioning, video understanding, speech recognition, and more. By leveraging multiple modalities, the model can exploit the strengths of each modality to improve overall performance.
- (3) Robustness to data limitations because multi-modal learning can mitigate the limitations of individual modalities by leveraging complementary information. If one modality lacks sufficient data or exhibits noise or ambiguity, the model can rely on other modalities to compensate for these shortcomings, resulting in improved robustness and generalization.
- (4) Richer context understanding, since combining different modalities allows for a more comprehensive understanding of context. For example, in natural language processing tasks, incorporating visual information alongside text can provide valuable visual context that enhances language understanding and generates more accurate responses.
- (5) Cross-modal transfer learning since multi-modal deep learning models can transfer knowledge between different modalities. Pretraining on one modality and fine-tuning on another can accelerate the learning process and improve performance, even with limited labeled data in the target modality.
- (6) Better human-like perception, since humans naturally integrate information from multiple senses to perceive and interpret the world. Multi-modal deep learning aims to mimic

this human-like perception by fusing information from diverse modalities, enabling machines to understand and interact with the environment in a more human-centric way.

- (7) Discovering hidden relationships because multi-modal learning can uncover hidden relationships and correlations between different modalities that may not be apparent in isolation. This can lead to new insights and discoveries, especially in domains where the data is inherently multi-modal, such as in healthcare, autonomous driving, and social media analysis.

These virtues make multi-modal deep learning a promising approach for a wide range of tasks and domains, allowing for richer and more nuanced data analysis, understanding, and decision-making and our findings provide further evidence of the competitiveness of multi-model deep learning models, particularly when leveraging more sophisticated architectures that incorporate late fusion strategies (Ramachandram and Taylor, 2017; Baltrušaitis et al., 2018), as seen in the extension of the model used by Montesinos-López et al. (2023). Additionally, our study benefits from the utilization of larger datasets.

The results of our study demonstrate the multi-modal DL models proposed outperform GBLUP models in certain traits and exhibit similar performance in others. However, when predicting for an entire year, the performance, while still comparable, is slightly reduced compared to the GBLUP model. This could be attributed to the relatively smaller training size available for the models in these scenarios, in which more exploration can be done where other strategy tuning parameters and loss function could be evaluated.

Our results agree with the growing evidence that multi-modal deep learning models are a powerful tool for predicting more efficiently in the context where multiple-inputs capture different portions of the signal of the response variable. Because the modelling process trains a particular deep neural network for

each input (modality), at the end, all the outputs of these deep neural networks are concatenated in a final deep neural network that produces the final predictions. The multi-modal deep learning for its architecture (Figure 1) facilitates the training process to efficiently capture the signal of the response and control of the overfitting. For these reasons, application of multi-modal deep learning models continues growing in many fields like health care, bioinformatics, computer vision, etc.

Finally, it is important to note that by leveraging the power of multi-modal deep learning, genomic prediction can benefit from the integration of diverse data sources, improved prediction accuracy, robustness to missing data, and enhanced interpretability, ultimately advancing our understanding of genetic traits and their implications in various applications, including precision medicine and agricultural breeding programs.

Conclusions

Using a moderately large dataset comprising 4464 lines evaluated for 5 agronomic traits under 3 or 4 different environments, we conducted a comparative analysis between GBLUP models implemented in the BGLR R package and a novel multi-modal deep learning (DL) model developed in this study. The results demonstrate the extended DL model presented achieved higher accuracy in predicting certain traits, specifically Maturity and Heading, when evaluated using the 5FCV. The DL model exhibited comparable accuracy to the GBLUP models for the remaining traits: Yield, Height, and Germination.

The DL approach utilized in this study extends and complements the previously proposed model, resulting in significant improvements in prediction accuracy for new environments. This finding further supports the notion that constructing individual networks for each modality and subsequently combining their outputs to feed into another network can yield more flexible and accurate models.

Data availability statement

Publicly available datasets were analyzed in this study. The phenotypic and genomic wheat data employed in this study can be downloaded from the following link <https://hdl.handle.net/11529/10548813> (Montesinos-López et al., 2023).

Author contributions

JC: Conceptualization, Investigation, Writing – review & editing. AM-L: Conceptualization, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. LC-H: Data curation, Project administration, Writing – review &

editing. SD: Investigation, Methodology, Writing – review & editing. GG: Conceptualization, Validation, Writing – review & editing. PV: Data curation, Visualization, Writing – review and editing. CS: Data curation, Funding acquisition, Supervision, Writing – review & editing. VG: Data curation, Validation, Writing – review & editing. ZT: Data curation, Investigation, Visualization, Writing – review & editing. MF: Software, Writing – review & editing. PP-R: Validation, Writing – review & editing. SR-P: Investigation, Software, Writing – review & editing. ML: Conceptualization, Methodology, Writing – review & editing. HL: Conceptualization, Methodology, Writing – review & editing. OM-L: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Open Access fees were received from the Bill & Melinda Gates Foundation. We acknowledge the financial support provided by the Bill & Melinda Gates Foundation (INV-003439 BMGF/FCDO Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG)) as well as the USAID projects (Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, Genes 2023, 14, 927 14 of 18 AGG-Maize Supplementary Project, AGG (Stress Tolerant Maize for Africa)) which generated the CIMMYT data analyzed in this study. We are also thankful for the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) through the Research Council of Norway for grants 301835 (Sustainable Management of Rust Diseases in Wheat) and 320090 (Phenotyping for Healthier and more Productive Wheat Crops). We acknowledge the support of the Window 1 and 2 funders to the Accelerated Breeding Initiative (ABI).

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.

References

- Balaji, T. K., Annavarapu, C. S. R., and Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Comput. Sci. Rev.* 40, 100395. doi: 10.1016/j.cosrev.2021.100395
- Baltrušaitis, T., Ahuja, C., and Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607
- Chandrasekaran, G., Nguyen, T. N., and Hemanth D, J. (2021). Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* 11, e1415. doi: 10.1002/widm.1415
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011
- Danilevicz, M. F., Bayer, P. E., Boussaid, F., Bennamoun, M., and Edwards, D. (2021). Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sens.* 13 (19), 3976. doi: 10.3390/rs13193976
- Duan, S., Shi, Q., and Wu, J. (2022). Multimodal sensors and ML-based data fusion for advanced robots. *Advanced Intelligent Syst.* 4, 2200213. doi: 10.1002/aisy.202200213
- FAO (2021). *Wheat* (Rome, Italy: Food and Agriculture Organization of the United Nations). Available at: <http://www.fao.org/faostat/en/#data/QC>.
- Garillos-Manlriquez, C. A., and Chiang, J. Y. (2021). Multimodal deep learning and visible-light and hyperspectral imaging for fruit maturity estimation. *Sensors* 21, 1288. doi: 10.3390/s21041288
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conf. Comput. Vision Pattern Recognition (CVPR)*, 770–778. doi: 10.1109/CVPR.2016.90
- Heffner, E. L., Jannink, J. L., and Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *Plant Genome* 4 (1), 65–75. doi: 10.3835/plantgenome.2010.12.0029
- Huang, S. C., Pareek, A., Zamanian, R., Banerjee, I., and Lungren, M. P. (2020). Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci. Rep.* 10 (1), 22147. doi: 10.1038/s41598-020-78888-w
- Jabeen, S., Li, X., Amin, M. S., Bourahla, O., Li, S., and Jabbar, A. (2023). A review on methods and applications in multimodal deep learning. *ACM Trans. Multimedia Computing Commun. Appl.* 19, 1–41. doi: 10.1145/3545572
- Jiang, Y., and Li, C. (2020). Convolutional neural networks for image-based high-throughput plant phenotyping: a review. *Plant Phenomics* 2020, 1–20. doi: 10.34133/2020/4152816
- Kick, D. R., Wallace, J. G., Schnable, J. C., Kolkman, J. M., Alaca, B., Beissinger, T. M., et al. (2023). Yield prediction through integration of genetic, environment, and management data through deep learning. *G3: Genes Genomes Genet.* 13, jkad006. doi: 10.1093/g3journal/jkad006
- Kline, A., Wang, H., Li, Y., Dennis, S., Hutch, M., Xu, Z., et al. (2022). Multimodal machine learning in precision health: a scoping review. *NPJ Digit. Med.* 5 (1), 171. doi: 10.1038/s41746-022-00712-8
- Liu, K., Li, Y., Xu, N., and Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. *arXiv*. [Preprint].
- Melotti, G., Premevida, C., and Gonçalves, N. (2020). “Multimodal deep-learning for object recognition combining camera and LIDAR data,” in *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. (Ponta Delgada, Portugal: IEEE), 177–182.
- Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Gianola, D., Hernández-Suárez, C. M., and Martín-Vallejo, J. (2018). Multi-trait, multi-environment deep learning modeling for genomic-enabled prediction of plant traits. *G3: Genes genomes Genet.* 8, 3829–3840. doi: 10.1534/g3.118.200728
- Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W., Fajardo-Flores, S. B., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22, 1–23. doi: 10.1186/s12864-020-07319-x
- Montesinos-López, A., Rivera, C., Pinto, F., Piñera, F., Gonzalez, D., Reynolds, M., et al. (2023). Multimodal deep learning methods enhance genomic prediction of wheat breeding. *G3: Genes Genomes Genet.* 13, jkad045. doi: 10.1093/g3journal/jkad045
- Morency, L. P., and Baltrušaitis, T. (2017). “Multimodal machine learning: integrating language, vision and speech,” in *Proceedings of the 55th annual meeting of the association for computational linguistics: Tutorial abstracts*. 3–5.
- Muroga, S., Miki, Y., and Hata, K. (2023). A comprehensive and versatile multimodal deep learning approach for predicting diverse properties of advanced materials. *arXiv*. [Preprint]. doi: 10.1002/adv.202302508
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). “Multimodal deep learning,” in *Proceedings of the 28th international conference on machine learning (ICML-11)*. 689–696.
- Ouyang, W., Chu, X., and Wang, X. (2014). “Multi-source deep learning for human pose estimation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2329–2336.
- Pérez-Rodríguez, P., and de los Campos, G. (2014). BGLR: a statistical package for whole genome regression and prediction. *Genetics* 198 (2), 483–495. doi: 10.1534/genetics.114.164442
- Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome* 5 (1), 103–113. doi: 10.3835/plantgenome.2012.06.0006
- Rahate, A., Walambe, R., Ramanna, S., and Kotecha, K. (2022). Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Inf. Fusion* 81, 203–239. doi: 10.1016/j.inffus.2021.12.003
- Ramachandram, D., and Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Magazine* 34, 96–108. doi: 10.1109/MSP.2017.2738401
- Rutkoski, J., Poland, J., Mondal, S., Autrique, E., Pérez, L. G., Crossa, J., et al. (2016). Canopy temperature and vegetation indices from high-throughput phenotyping improve accuracy of pedigree and genomic selection for grain yield in wheat. *G3: Genes Genomes Genet.* 6, 2799–2808. doi: 10.1534/g3.116.032888
- Srivastava, N., and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. *Adv. Neural Inf. Process. Syst.* 25, 1–9.
- Stahlschmidt, S. R., Ulfenborg, B., and Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. *Brief. Bioinform.* 23 (2), bbab569. doi: 10.1093/bib/bbab569
- Summaira, J., Li, X., Shoib, A. M., Li, S., and Abdul, J. (2021). Recent advances and trends in multimodal deep learning: a review. *arXiv*. [Preprint].
- Venugopalan, J., Tong, L., Hassanzadeh, H. R., and Wang, M. D. (20213254). Multimodal deep learning models for early detection of alzheimer’s disease stage. *Sci. Rep.* 11 (1). doi: 10.1038/s41598-020-74399-w
- Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L. P. (2018). “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.
- Zhou, J., Li, J., Wang, C., Wu, H., Zhao, C., and Teng, G. (2021). Crop disease identification and interpretation method based on multimodal deep learning. *Comput. Electron. Agric.* 189, 106408. doi: 10.1016/j.compag.2021.106408



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Tingxi Yu,
Chinese Academy of Agricultural Sciences
(CAAS), China
João Ricardo Bachega Feijó Rosa,
RB Genetics & Statistics Consulting, Brazil

*CORRESPONDENCE

Abelardo Montesinos-López
✉ abelardo.montesinos0233@
academicos.udg.mx
José Crossa
✉ j.crossa@cgiar.org

RECEIVED 04 December 2023

ACCEPTED 11 April 2024

PUBLISHED 15 May 2024


CITATION

Montesinos-López OA, Crespo-Herrera L,
Pierre CS, Cano-Paez B, Huerta-Prado GI,
Mosqueda-González BA, Ramos-Pulido S,
Gerard G, Alnowibet K, Fritsche-Neto R,
Montesinos-López A and Crossa J (2024)
Feature engineering of environmental
covariates improves plant
genomic-enabled prediction.
Front. Plant Sci. 15:1349569.
doi: 10.3389/fpls.2024.1349569

COPYRIGHT

© 2024 Montesinos-López, Crespo-Herrera,
Pierre, Cano-Paez, Huerta-Prado, Mosqueda-
González, Ramos-Pulido, Gerard, Alnowibet,
Fritsche-Neto, Montesinos-López and Crossa.
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.

Feature engineering of environmental covariates improves plant genomic-enabled prediction

Oswal A. Montesinos-López¹, Leonardo Crespo-Herrera²,
Carolina Saint Pierre², Bernabe Cano-Paez³,
Gloria Isabel Huerta-Prado⁴,
Brandon Alejandro Mosqueda-González⁵, Sofia Ramos-Pulido⁶,
Guillermo Gerard², Khalid Alnowibet⁷,
Roberto Fritsche-Neto⁸, Abelardo Montesinos-López^{6*}
and José Crossa ^{2,8,9,10*}

¹Facultad de Telemática, Universidad de Colima, Colima, Mexico, ²International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Edo. de Mexico, Mexico, ³Facultad de Ciencias, Universidad Nacional Autónoma de México (UNAM), México City, Mexico, ⁴Independent consultant, Zinacatepec, Puebla, Mexico, ⁵Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), México City, Mexico, ⁶Centro Universitario de Ciencias Exactas e Ingenierías (CUCEI), Universidad de Guadalajara, Guadalajara, Jalisco, Mexico, ⁷Department of Statistics and Operations Research, King Saud University, Riyadh, Saudi Arabia, ⁸Louisiana State University, Baton Rouge, LA, United States, ⁹Distinguished Scientist Fellowship Program, King Saud University, Riyadh, Saudi Arabia, ¹⁰Instituto de Socioeconomía, Estadística e Informática, Colegio de Postgraduados, Montecillos, Edo. de México, Texcoco, Mexico

Introduction: Because Genomic selection (GS) is a predictive methodology, it needs to guarantee high-prediction accuracies for practical implementations. However, since many factors affect the prediction performance of this methodology, its practical implementation still needs to be improved in many breeding programs. For this reason, many strategies have been explored to improve the prediction performance of this methodology.

Methods: When environmental covariates are incorporated as inputs in the genomic prediction models, this information only sometimes helps increase prediction performance. For this reason, this investigation explores the use of feature engineering on the environmental covariates to enhance the prediction performance of genomic prediction models.

Results and discussion: We found that across data sets, feature engineering helps reduce prediction error regarding only the inclusion of the environmental covariates without feature engineering by 761.625% across predictors. These results are very promising regarding the potential of feature engineering to enhance prediction accuracy. However, since a significant gain in prediction accuracy was observed in only some data sets, further research is required to guarantee a robust feature engineering strategy to incorporate the environmental covariates.

KEYWORDS

genomic selection, plant breeding, environmental covariates, feature engineering, feature selection

Introduction

The global population's rapid growth is increasing food demand, but climate change impacts crop productivity. Plant breeding is essential for high-yield, quality cultivars. Wheat production soared from 200 million tons in 1961 to 775 million tons in 2023 without expanding cultivation, thanks to improved cultivars and agricultural practices (FAO, 2023). Traditional methods used pedigree and observable traits, but DNA sequencing introduced genomic insights. Genomic selection (GS) relies on DNA markers, offering advantages over traditional methods (Crossa et al., 2017).

Numerous studies have investigated the efficacy of GS compared to traditional phenotypic selection across various crops and livestock. Butoto et al. (2022) observed that both GS and phenotypic selection were equally effective in enhancing resistance to Fusarium ear rot and reducing feminizing contamination in maize. Similarly, Sallam and Smith (2016) demonstrated that integrating GS into barley breeding programs targeting yield and Fusarium head blight (FHB) resistance yielded comparable gains in selection response to traditional phenotypic methods. Moreover, GS offered the added benefits of shorter breeding cycles and reduced costs. In contrast, research in maize breeding conducted by Beyene et al. (2015) and Gesteiro et al. (2023) revealed that GS outperformed phenotypic selection, resulting in superior genetic gains. These comparative findings underscore the considerable advantages of GS in optimizing breeding outcomes across diverse agricultural settings.

GS revolutionizes plant and animal breeding by leveraging high-density markers across the genome. It operates on the principle that at least one genetic marker is in linkage disequilibrium with a causative QTL (Quantitative Trait Locus) for the desired trait (Meuwissen et al., 2001). This method transforms breeding in several ways: a) Identifying promising genotypes before planting; b) Improving precision in selecting superior individuals; c) Saving resources by reducing extensive phenotyping; d) Accelerating variety development by shortening breeding cycles; e) Intensifying selection efforts; f) Facilitating the

selection of traits difficult to measure; g) Enhancing the accuracy of the selection process (Bernardo and Yu, 2007; Heffner et al., 2009; Desta and Ortiz, 2014; Abed et al., 2018; Budhlakoti et al., 2022).

The GS methodology, embraced widely, expedites genetic improvements in plant breeding programs (Desta and Ortiz, 2014; Bassi et al., 2016; Xu et al., 2020). Utilizing advanced statistical and machine learning models (Montesinos-López et al., 2022), GS efficiently selects individuals within breeding populations. Deep learning, a subset of machine learning, has also shown promise in GS (Montesinos-López et al., 2021; Wang et al., 2023). This selection process relies on data from a training population, encompassing both phenotypic and genotypic information (Crossa et al., 2017).

The Deep Neural Network Genomic Prediction (DNNGP) method of Wang et al. (2023) represents a novel advanced on deep-learning genomic predictive approach. The authors compared the DNNGP with other genomic prediction methods for various traits using genotypic and transcriptomics on maize data. They demonstrated that DNNGP outperformed GBLUP in most datasets. For instance, for maize days to anthesis (DTA) trait, DNNGP showed superiority over GBLUP by 619.840% and 16.420% using gene expression and Single Nucleotide Polymorphism (SNP) data, respectively. When utilizing genotypic data, DNNGP achieved a prediction accuracy of 0.720 for DTA, while GBLUP reached 0.580. However, the study found varied patterns in prediction accuracy for other traits.

Following rigorous training, these models utilize genotypic data to predict breeding or phenotypic values for traits within a target population (Budhlakoti et al., 2022). The GS methodology is versatile, accommodating various scenarios including multi-trait considerations (Calus and Veerkamp, 2011), known major genes and marker-trait associations, Genotype × Environment interaction (GE) (Crossa et al., 2017), and integration of other omics data (Hu et al., 2021; Wu et al., 2022) such as transcriptomics, metabolomics, and proteomics. GE influences phenotypic trait values across diverse environments, underscoring its importance in association and prediction models. Jarquin et al. (2014) introduced a framework significantly improving prediction accuracy in the

presence of GE, yet without considering environmental covariates. To enhance accuracy further, recent studies are integrating environmental information into genomic prediction models.

Jarquín et al. (2014) framework lacks consideration of environmental covariates, prompting recent studies to integrate such information to enhance prediction accuracy. For instance, Montesinos-López et al. (2023) and Costa-Neto et al. (2021a, 2021b) demonstrated significant improvements. Conversely, studies by Monteverde et al. (2019); Jarquín et al. (2020), and Rogers et al. (2021) showed modest or negligible enhancements, revealing the ongoing challenge of effectively integrating environmental data into genomic prediction models.

Achieving high prediction accuracy in GS faces significant challenges due to genetic complexities, environmental variations, and data constraints (Juliana et al., 2018). Complex traits involve multiple gene influences, while environmental conditions can alter trait expression (Desta and Ortiz, 2014; Crossa et al., 2017). Phenotyping and marker data quality are critical, and issues like overfitting and population structure can compromise prediction precision (Budhlakoti et al., 2022). Ongoing research focuses on improving models, increasing marker density, and enhancing data quality to refine genomic prediction accuracy (Crossa et al., 2017; Budhlakoti et al., 2022).

Ongoing efforts focus on refining GS accuracy through various optimizations. This includes fine-tuning training and testing sets for improved precision (Rincént et al., 2012; Akdemir et al., 2015). Researchers are also evaluating diverse statistical machine learning methods to develop robust models with minimal fine-tuning yet high accuracy (Montesinos-López et al., 2022). Moreover, integrating additional omics data, such as phenomics and transcriptomics, aims to bolster GS accuracy and identify potent predictors for target traits (Montesinos-López et al., 2017; Krause et al., 2019; Monteverde et al., 2019; Hu et al., 2021; Costa-Neto et al., 2021a, b; Rogers and Holland, 2022; Wu et al., 2022). These endeavors seek to enhance GS predictive capabilities by leveraging diverse information sources.

Feature engineering (FE) is crucial in improving machine learning model performance by selecting, modifying, or creating new features from raw data. It transforms input data into a more representative and informative format, capturing relevant patterns and relationships, and enhancing the model's generalization ability. FE involves various tasks like selecting optimal features, generating new features, normalization/scaling, handling missing values, and encoding categorical variables. For instance, techniques like Principal Component Analysis (PCA) can transform correlated features into uncorrelated ones (Lam et al., 2017; Dong and Liu, 2018; Khurana et al., 2018). FE's popularity is rising due to its ability to enhance model performance, extract meaningful information from complex data, improve interpretability, and boost efficiency. Successful implementations include sentiment analysis, image recognition, and predictive maintenance, showcasing FE's effectiveness across domains (Nargesian et al., 2017; Carrillo-de-Albornoz et al., 2018; Yurek and Birant, 2019). In genomic prediction, FE has also been successful, as demonstrated by Bermingham et al. (2015) and Afshar and Usefi (2020). These

examples underscore FE critical role in various domains, leading to more accurate machine learning applications (Dong and Liu, 2018).

The impact of feature engineering (FE) on reducing prediction error varies depending on the dataset, problem, and quality of FE. Well-crafted features can notably minimize prediction error in some cases, but the exact improvement is context-specific and not guaranteed. Effective FE can enhance model performance significantly, albeit its extent varies case by case (Heaton, 2016; Dong and Liu, 2018).

To optimize genomic selection's predictive accuracy, it's vital to adopt innovative methodologies that account for its multifaceted influences. FE in genomic prediction offers a promising approach by enhancing prediction quality, uncovering genetic insights, customizing models to specific needs, improving interpretability, and minimizing data noise. In this paper, we investigate FE applied to environmental covariates to assess its potential in enhancing prediction performance within the context of genomic selection.

Materials and methods

Dataset USP

The University of São Paulo (USP) Maize, *Zea mays* L., dataset is sourced from germplasm developed by the Luiz de Queiroz College of Agriculture at the University of São Paulo, Brazil. An experiment was conducted between 2016 and 2017 involving 49 inbred lines, yielding a total of 906 F1 hybrids, of which 570 were assessed across eight diverse environments for grain yield (GY). These environments were created by combining two locations, two years, and two nitrogen levels. However, we specifically used data from four distinct environments for this research, each containing 100 hybrids. It's important to note that these environments had varying soil types and climatic conditions, and the study integrated data from 248 covariates related to these environmental factors. The parent lines underwent genotyping through the Affymetrix Axiom Maize Genotyping Array, resulting in a dataset of 54,113 high-quality SNPs after applying stringent quality control procedures. Please refer to Costa-Neto et al. (2021a) for further comprehensive information on this dataset.

Dataset Japonica

The Japonica dataset comprises 320 rice (*Oryza sativa* L.) genotypes drawn from the Japonica tropical rice population. This dataset underwent evaluations for the same four traits (GY, PHR: percentage of head rice, GC: percentage of chalky grains, PH: plant height) as the Indica population, but in this case, it was conducted across five distinct environments spanning from 2009 to 2013. Covariates were meticulously measured three times a year, covering three developmental stages (maturation, reproductive, and vegetative). This dataset comprises a non-balanced set of 1,051 assessments recorded across these five diverse environments. Additionally, each genotype within this dataset was meticulously

evaluated for 16,383 SNP markers that remained after rigorous quality control procedures, with each marker being represented as 0, 1, or 2. For more comprehensive information on this dataset, please refer to [Monteverde et al. \(2019\)](#).

Dataset G2F

These three distinct datasets correspond to the Maize Crop, *Zea mays* L., for years 2014 (G2F_2014), 2015 (G2F_2015), and 2016 (G2F_2016) from the Genomes to Fields maize project ([Lawrence-Dill, 2017](#)), as outlined by [Rogers and Holland \(2022\)](#). These datasets collectively encompass a wealth of phenotypic, genotypic, and environmental information. To narrow the focus, our analysis primarily includes four specific traits: Grain_Moisture_BLUE (GM_BLUE), Grain_Moisture_weight (GM_Weight), Yield_Mg_ha_BLUE (YM_BLUE), and Yield_Mg_ha_weight (YM_Weight), carefully selected from a larger pool of traits detailed by [Rogers and Holland \(2022\)](#). Across these three years, the study involves 18, 12, and 18 distinct environments for the years 2014 (G2F_2014), 2015 (G2F_2015) and 2016 (G2F_2016), respectively. Regarding genotype numbers, the dataset for 2014 consisted of 781 genotypes, the dataset for 2015 featured 1,011 genotypes, and the dataset for 2016 comprised 456 genotypes. The analysis relies on 20,373 SNP markers that have already undergone imputation and filtering, following the methodology outlined by [Rogers et al. \(2021\)](#) and [Rogers and Holland \(2022\)](#). Additive allele calls are documented as minor allele counts, represented as 0, 1, or 2. For more detailed insights into these datasets, we recommend consulting the comprehensive description provided in [Lawrence-Dill \(2017\)](#) and [Rogers and Holland \(2022\)](#).

It is worth noting that each data set presents unique sets of environments. However, concerning traits, the G2F_2014, G2F_2015, and G2F_2016 datasets share identical traits, as do the Japonica dataset.

Statistical models

The four predictors under a genomic best linear unbiased predictor (GBLUP; [Habier et al., 2007](#); [VanRaden, 2008](#)) model are described below.

Predictor P1: E+G

This predictor is represented as

$$Y_{ij} = \mu + E_i + g_j + \epsilon_{ij}, \quad (1)$$

where Y_{ij} denotes the response variable in environment i and genotype j . μ denotes the population mean; E_i are the random effects of environments, g_j , $j = 1, \dots, J$, denotes the random effects of lines, and ϵ_{ij} denotes the random error components in the model assumed to be independent normal random variables with mean 0 and variance σ^2 . In the context of this predictor E+G, X , denotes the matrix of markers and M the matrix of centered and standardized markers. Then $G = \frac{MM^T}{p}$ ([VanRaden, 2008](#)), where p

is the number of markers. Z_g is the design matrix of genotypes (lines) of order $n \times J$, G is the genomic relationship-matrix computed using markers ([VanRaden, 2008](#)). Therefore, the random effect of lines is distributed as $g = (g_1, \dots, g_J)^T \sim N_J(0, \sigma_g^2 Z_g G Z_g^T)$. This model (1) was implemented in the BGLR library of [Pérez and de los Campos \(2014\)](#). Therefore, the linear kernel matrix for the genotype effect was determined by calculating the “covariance” structure of the genotype predictor ($Z_g g$) as $K_g = Z_g G Z_g^T$.

On the other hand, the linear kernel matrix for the Environment effect was computed using three different techniques: not using environmental covariates (NoEC), with environmental covariates (EC), and with environmental covariates with FE.

- **NoEC:** Under this NoEC technique, the resulting linear kernel of environments was computed as $K_E = X_E X_E^T / I$, where I denotes the number of environments and X_E the design matrix of environments with zeros and ones, with ones in positions of specific environments.
- **EC:** The EC technique involved selecting and scaling the environmental covariates (EC) that exhibited a relevant Pearson’s correlation with the response variable. Covariates are selected if their Pearson’s correlation with the response variable exceeds 0.5 in each training set per trait. Notably, covariate selection excludes response variables in the testing set, representing the environment to predict. Covariates meeting a correlation of at least 0.5 are used; otherwise, lower thresholds like 0.3 or 0.4 are considered. Correlations below these values indicate training without environmental covariates.
- The resulting set of selected EC’s was then used to compute an environmental linear kernel, denoted as K_{EC} of order $I \times I$. After using this kernel, the expanded environmental kernel was computed as $K_{EC} = X_E K_{EC} X_E^T / I$, which was used in the Bayesian model. The scaling of each environmental covariate was done by subtracting its respective mean and dividing by its corresponding standard deviation.
- **FE:** The Feature Engineering (FE) technique involved computing various mathematical transformations between all possible pairs of ECs, including addition, difference, product, and ratio, as well as other commonly used transformations such as inverses, square powers, root squares, logarithms, and some Box-Cox transformations for each EC. These transformations were used to generate new variables through FE. The transformation of addition, difference, product and ratio were implemented for each pair of environmental covariates, that is, there were built a total the n_{cov} choose two new covariates, with n_{cov} denoting the number of environmental covariates in each data set. While with transformations such as inverses ($1/x$), square powers (x^2), root squares (\sqrt{x}), natural logarithms [$\ln(x)$], and Box-Cox transformations for each environmental covariate was created only one new environmental covariate. Then the original and new

environmental covariates were concatenated in a matrix and then were submitted to the selection process explained above. Then under the FE approach these resulting covariates are used to compute the new environmental kernel matrix (K_{EFE}).

Predictor P2: E+G+GE

The E+G+GE predictor is similar to P1 (Equation 1) but also accounts for the differential response of cultivars in environments, that is GE. This is achieved by taking the product of the kernel matrices of the genotype (G) and environment (E) predictors, that is, they were computed as $K_g \circ K_{E_{NoEC}}$ (for NoEC), $K_g \circ K_{E_{EC}}$ (for EC) or $K_g \circ K_{E_{FE}}$ (for FE), which serves as the kernel matrix for the GE. In general, adding the GE interaction to the statistical machine learning model increases the genomic prediction accuracy (Jarquin et al., 2014; Crossa et al., 2017). Also, it is important to point out that under this predictor (P2) variance components and heritability of each trait in each data set were obtained under a Bayesian framework using the complete data set (i.e., no missing values allowed). For this computation all the terms were entered as random effects into the model but without taking into account the environmental covariates.

Predictor P3: E+G+BRR

The E+G+BRR predictor is similar to P1 (Equation 1), but incorporating the ECs as fixed effects in a Bayesian Ridge Regression (BRR) framework, that is, regression coefficients are assigned normal independent and identically distributed normal distributions, with mean zero and variance σ_β^2 . See details of BRR in Pérez and de los Campos (2014).

Predictor P4: E+G+GE+BRR

The E+G+GE+BRR predictor is similar to P2, but also incorporates ECs as fixed effects in a Bayesian Ridge Regression (BRR) framework (see Appendix for brief details on Bayesian Ridge Regression). The priors used for GBLUP and BRR in BGLR are those default settings which are given with details in Pérez and de los Campos (2014). In this study, we found these default settings to be suitable, as we experimented with various configurations of the prior hyperparameters for the GBLUP and BRR models on the USP and G2F_2014 datasets. Remarkably, all configurations yielded identical predictions. Consequently, for the remaining datasets, we opted to utilize only the default settings.

Evaluation of prediction performance

The cross-validation approach used in this study involved leaving one environment out. In each iteration, the data from a single-environment served as the testing set, while the data from all other families constituted the training set (Montesinos-López et al., 2022). The number of iterations was equal to the number of environments to ensure that each environment was used as the testing set exactly one time. This method was employed to assess the model's ability to predict information from a complete environment using data from other environments.

To evaluate the predictive performance we used the Mean Square Error (MSE) that quantifies the prediction error by measuring the squared deviation between observed and predicted values on the testing set. The MSE was computed for each scenario evaluated (NoEC, EC and FE) and then for comparing these three scenarios was computed the relative efficiencies as:

$$RE_{NoEC_vs_EC} = \left(\frac{MSE(NoEC)}{MSE(EC)} \right)$$

$$RE_{NoEC_vs_FE} = \left(\frac{MSE(NoEC)}{MSE(FE)} \right)$$

$$RE_{EC_vs_FE} = \left(\frac{MSE(EC)}{MSE(FE)} \right)$$

$RE_{NoEC_vs_EC}$ compares the prediction performance of EC vs NoEC, $RE_{NoEC_vs_FE}$ compares the prediction performance of FE vs NoEC and $RE_{EC_vs_FE}$ compares the prediction performance of FE vs EC. When $RE_{NoEC_vs_EC} > 1$ the best prediction performance was obtained by the EC strategy, while when $RE_{NoEC_vs_EC} < 1$ the strategy NoEC was the best. While when the relative efficiencies are equal to 1 means that both methods had equal prediction performance. The same interpretation applies for the other comparisons in terms of RE.

Results

The results are given in three sections for three datasets (Japonica, USP and G2F_2016). For each section we provided the results for the four predictor models under study (E+G, E+G+GE, E+G+BRR, E+G+GE+BRR) and under each predictor we compared three strategies for the use of the environmental covariates: NoEC, using environmental covariables (EC) and using environmental covariables with FE. Additionally, Appendix A contains comprehensive details of the BRR model utilized in this study. Furthermore, Appendix B offers extensive information on the outcomes for Japonica, USP, and G2F_2016 datasets, which are outlined in Table B1–Table B2, Table B3, Table B4, Table B4–Table B5 respectively. Additionally, Table B7 in this appendix presents the variance components and heritability of each trait within every dataset. For the results pertaining to datasets G2F_2014 and G2F_2015, please refer to the Supplementary Materials section.

Japonica dataset

Predictor: E+G

Figure 1A provides a summary of Table B1 across traits and reveals that FE outperformed EC in most environments with improvements of 20.260% (2010), 38.920% (2011), 1.750% (2012), and 25.470% (2013). This results in an average RE of 1.1567. EC, on the other hand, outperformed NoEC in most environments with improvements of 121.200% (2009), 48.080% (2010), and 8.140% (2012), resulting in an average RE of 1.277. Likewise, FE outperformed NoEC in 101.240% (2009), 59.560% (2010), and

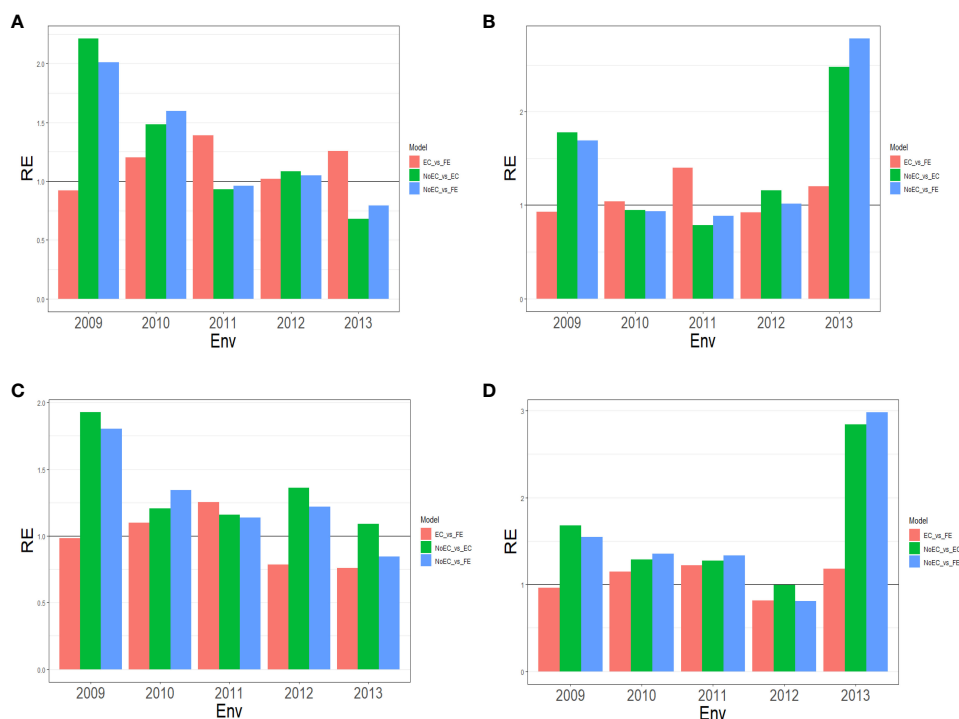


FIGURE 1

The three relative efficiencies, considering EC_vs_FE, NoEC_vs_EC, and NoEC_vs_FE, for Japonica dataset, for predictors (A) E+G, (B) E+G+GE, (C) E+G+BRR and (D) E+G+GE+BRR in terms of mean squared error (MSE) for each Environment across traits.

4.710% (2012), with slight losses in other environments, but an average RE of 1.2814. This indicates that using EC and FE surpassed NoEC by 27.730% and 28.140%, respectively. These calculations are derived from the results presented in Table B1.

Predictor: E+G+GE

Figure 1B summarizes the findings from Table B1 across traits, illustrating the comparative performance of FE, EC, and NoEC techniques in various environments. The results indicate that FE outperformed EC in the majority of environments, with improvements of 4.280% (2010), 40.050% (2011), and 20.220% (2013), resulting in an average RE of 1.099. On the other hand, EC outperformed NoEC in most environments, with improvements of 78.070% (2009), 16.100% (2012), and 147.980% (2013), yielding an average RE of 1.430. Furthermore, FE surpassed the conventional NoEC technique by 68.990% (2009), 1.780% (2012), and 178.280% (2013), with an average RE of 1.462. These results indicate that using EC and FE techniques outperformed the conventional NoEC technique by 43.040% and 46.150%, respectively. The calculations are derived from the outcomes presented in Table B1.

Predictor: E+G+BRR

Figure 1C provides an overview of Table B2 across traits. It reveals that FE outperformed EC only in environments 2010

(9.630%) and 2011 (25.340%), resulting in an average RE of 0.975. On the other hand, EC outperformed NoEC in all environments, with percentages of improvement of 92.640% (2009), 20.690% (2010), 15.960% (2011), 36.170% (2012), and 9.070% (2013), and an average RE of 1.349. Additionally, FE outperformed the NoEC technique in 80.390% (2009), 34.120% (2010), 13.690% (2011), and 21.950% (2012) of the environments with a slight loss in 2013, but an average RE of 1.269. These findings indicate that using EC and FE techniques surpassed NoEC in 34.910% and 26.940% of the environments, respectively. The calculations are based on the results presented in Table B2.

Predictor: E+G+GE+BRR

Figure 1D summarizes the findings from Table B2 across traits. It reveals that FE displayed a superior performance over EC in environments 2010 (14.770%), 2011 (21.700%), and 2013 (17.870%), resulting in an average RE of 1.064. On the other hand, EC outperformed NoEC in most environments, namely 67.750% (2009), 28.390% (2010), 27.210% (2011), and 183.970% (2013), with an average RE of 1.614. Moreover, FE outperformed NoEC in most environments, specifically 54.260% (2009), 35.520% (2010), 33.140% (2011), and 197.980% (2013), with an average RE of 1.604. These findings indicate that using EC and FE surpassed NoEC in 61.390% and 60.460% of cases, respectively. The computations for these results were based on the findings presented in Table B2.

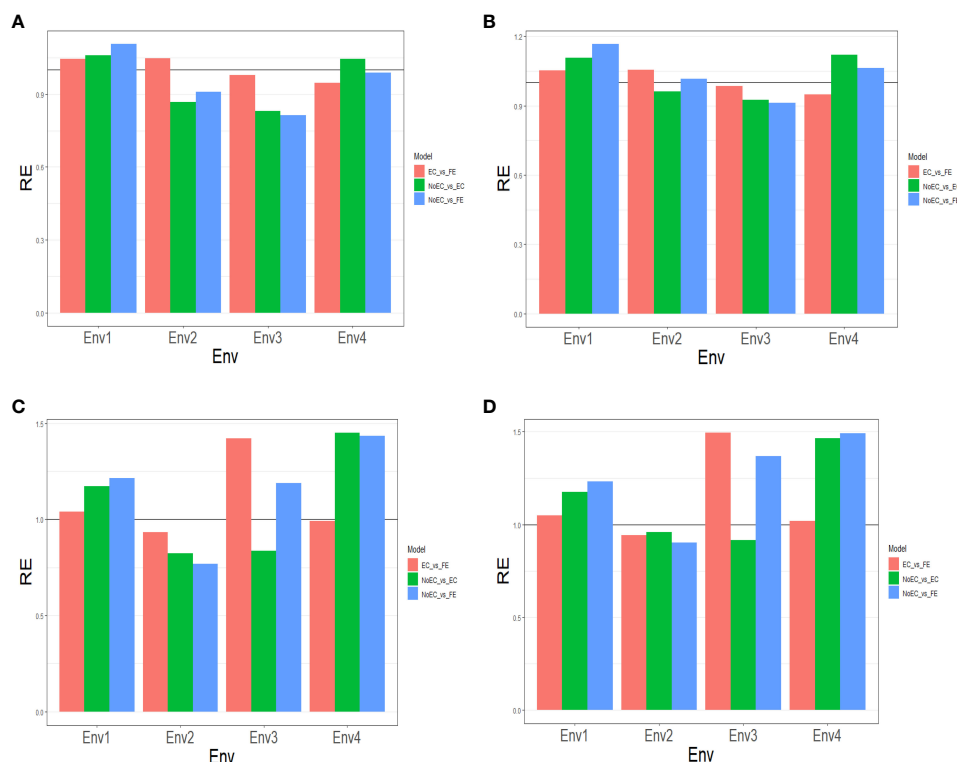


FIGURE 2

The three relative efficiencies, considering EC_vs_FE, NoEC_vs_EC, and NoEC_vs_FE, for USP dataset, for predictors (A) E+G, (B) E+G+GE, (C) E+G+BRR and (D) E+G+GE+BRR in terms of mean squared error (MSE) for each Environment.

USP dataset

Predictor: E+G

Figure 2A and Table B3 provide the results of our comparison between the NoEC and FE techniques using the RE metric. FE outperformed the NoEC technique only in Env1 (1.107), displaying an improvement of 10.670%. However, in Env2 (0.910), Env3 (0.8123), and Env4 (0.989), the NoEC technique surpassed FE, resulting in an average RE of 0.955. This average RE indicates a general loss of 4.520% when using FE compared to NoEC (see Table B3).

Predictor: E+G+GE

Figure 2B and Table B3 provide the results of our comparison between the NoEC and FE techniques based on the RE metric, including the fact that the use of FE outperformed the use of NoEC in environments Env1 (1.167), Env2 (1.016), and Env4 (1.064), resulting in respective improvements of 16.670%, 1.550%, and 6.390%. However, in Env3 (0.912), the NoEC technique outperformed FE, resulting in an average RE of 1.040. This average RE indicates a general improvement of 4.000% of the FE technique regarding the NoEC method. For more detailed information, see Table B3.

Predictor: E+G+BRR

Based on Figure 2C and Table B4, our comparison between the NoEC and FE techniques using the RE metric reveals that FE

outperformed the NoEC technique in environments Env1 (1.216), Env3 (1.189), and Env4 (1.435), displaying improvements of 21.580%, 18.890%, and 43.500%, respectively. However, in Env2 (0.768), the NoEC technique outperformed using FE. In general, FE outperformed NoEC by 15.200% since an average RE of 1.152 was observed (see Table B4).

Predictor: E+G+GE+BRR

Finally, based on the analysis presented in Figure 2D and Table B4, we compared the NoEC and FE techniques using the RE metric. The results indicate that FE outperformed NoEC in Env1 (1.231), Env3 (1.368), and Env4 (1.491), displaying improvements of 23.090%, 36.760%, and 49.080%, respectively. However, in Env2 (0.901), the NoEC technique outperformed FE, although, FE outperformed the NoEC in general terms, since an average RE of 1.248 was observed (see Table B4).

G2F_2016 dataset

Predictor: E+G

Figure 3A summarizes Table B5 across different environments for each trait. It reveals that FE outperformed EC in all traits, achieving improvements of 87.970% (Grain_Moisture_BLUE), 58.100% (Grain_Moisture_weight), 21.030% (Yield_Mg_ha_BLUE), and 89.600% (Yield_Mg_ha_weight), resulting in an average RE of 1.642. In contrast, EC outperformed NoEC in most traits, with

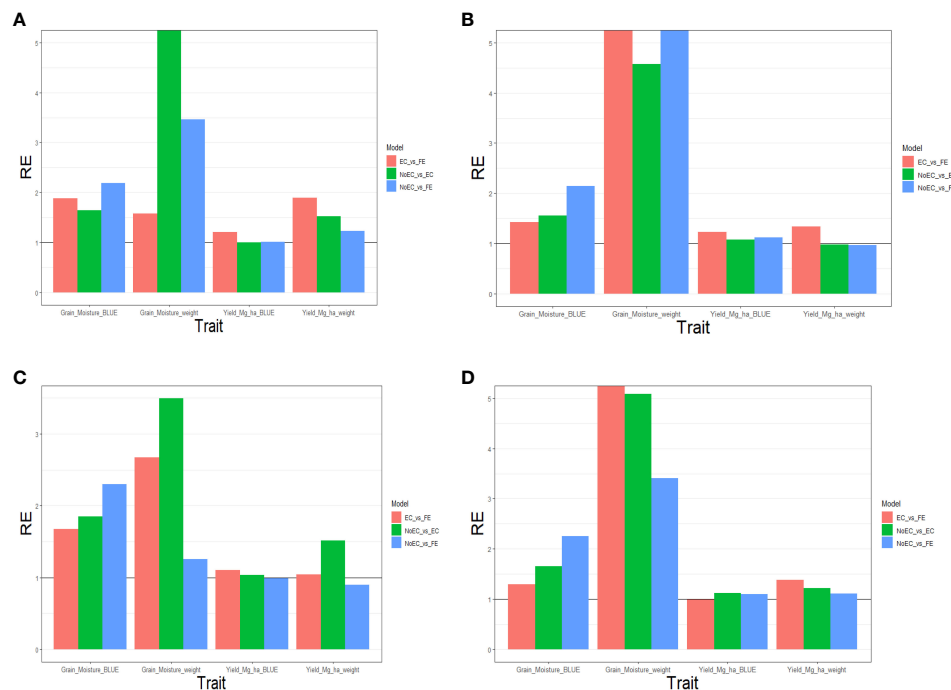


FIGURE 3

The three relative efficiencies, considering EC_vs_FE, NoEC_vs_EC, and NoEC_vs_FE, for G2F_2016 dataset, for predictors (A) E+G, (B) E+G+GE, (C) E+G+BRR and (D) E+G+GE+BRR in terms of mean squared error (MSE) for each trait across environments.

improvements of 63.960% (Grain_Moisture_BLUE), 1682.340% (Grain_Moisture_weight), and 52.860% (Yield_Mg_ha_weight), yielding an average RE of 5.497. Additionally, FE surpassed NoEC in all traits, with enhancements of 119.370% (Grain_Moisture_BLUE), 245.980% (Grain_Moisture_weight), 1.400% (Yield_Mg_ha_BLUE), and 22.630% (Yield_Mg_ha_weight), resulting in an average RE of 1.974. These findings indicate that both EC and FE techniques outperformed NoEC by 449.740% and 97.350%, respectively. The computations are based on the results presented in Table 5B.

Predictor: E+G+GE

Figure 3B and Table B5 shows that for the Yield_Mg_ha_weight trait, the NoEC technique achieved the best performance in most environments, as shown by the MSE values (DEH1_2016 [0.051], GAH1_2016 [0.026], IAH1_2016 [2.914], IAH2_2016 [0.069], MIH1_2016 [0.055], MNH1_2016 [0.146], NEH1_2016 [0.033], NYH2_2016 [0.449] and OHH1_2016 [1.202]). On average, there were slight losses of 2.210% and 2.570% when comparing EC versus NoEC and FE versus NoEC, respectively. This suggests that EC and FE techniques could have performed more adequately than the conventional NoEC technique. However, comparing EC and FE techniques based on RE showed that FE outperformed EC in most environments under NoEC, resulting in an average RE of 1.339, indicating a superiority of 33.930% for FE (see Table 5B).

Predictor: E+G+BRR

Figure 3C summarizes the findings from Table B6 across environments for each trait. It shows that FE outperformed EC in all characteristics, with improvements of 67.090% (Grain_Moisture_BLUE),

167.270% (Grain_Moisture_weight), 10.650% (Yield_Mg_ha_BLUE), and 3.960% (Yield_Mg_ha_weight), resulting in an average RE of 1.622. Additionally, EC outperformed NoEC in all traits, with improvements of 84.880% (Grain_Moisture_BLUE), 249.510% (Grain_Moisture_weight), 3.780% (Yield_Mg_ha_BLUE), and 51.630% (Grain_Moisture_weight), resulting in an average RE of 1.975. Furthermore, FE outperformed NoEC only in the traits Grain_Moisture_BLUE (129.850%) and Grain_Moisture_weight (25.410%), with an average RE of 1.360. These results indicate that EC and FE techniques outperformed the conventional NoEC technique in 62.240% and 36.020% of cases, respectively. These calculations are derived from the results presented in Table B6.

Predictor: E+G+GE+BRR

Figure 3D summarizes the results from Table B6 across different traits. It shows that FE outperformed EC in the majority of traits, specifically by 29.090% for Grain_Moisture_BLUE, 689.960% for Grain_Moisture_weight, and 38.420% for Yield_Mg_ha_weight. This leads to an average RE of 2.893. On the other hand, EC outperformed NoEC in all traits, with improvements of 65.180% for Grain_Moisture_BLUE, 408.510% for Grain_Moisture_weight, 11.690% for Yield_Mg_ha_BLUE, and 22.200% for Yield_Mg_ha_weight. The average RE for EC compared to NoEC is 2.269. Furthermore, FE outperformed NoEC in all traits, with improvements of 125.150% for Grain_Moisture_BLUE, 240.900% for Grain_Moisture_weight, 9.490% for Yield_Mg_ha_BLUE, and 11.380% for Yield_Mg_ha_weight. The average RE for FE compared to NoEC is 1.967. These results indicate that using EC and FE outperformed NoEC by 126.890% and 96.730%,

TABLE 1 Summary of relative efficiencies (RE) across data sets for each predictor.

Predictor	NoEC_vs_EC_	EC_vs_FE	NoEC_vs_FE
E+G	2.573	8.419	2.131
E+G+BRR	1.614	6.574	2.641
E+G+GE	2.489	4.473	3.141
E+G+GE+BRR	4.882	12.138	7.692
Average	2.889	7.901	3.901

NoEC_vs_EC denotes the RE of no using environmental covariates (NoEC) vs using environmental covariates (EC), EC_vs_FE denotes the RE efficiency of comparing using EC vs using the environmental covariates with feature engineering (FE) and NoEC_vs_FE is the RE of using FE regarding of no using environmental covariates (NoEC).

respectively. These computations are derived from the outcomes of Table B6.

Summary across data sets for each predictor

In Table 1 we can observe that in any of the four predictors using environmental covariates improve prediction accuracy at least 61.400% regarding of not using the environmental covariates (NoEC_vs_EC). Also, we can see in this same table that using FE improves the prediction performance in the four predictors regarding of using the original environmental covariates (EC_vs_FE) in at least 347.300%. Regarding using FE and not using environmental covariates (NoEC_vs_FE) we can observe that also in the four predictors using FE outperform by at least 113.100% not using the environmental covariates. Also, we observed that in many cases adding directly the environmental covariates (EC) not improve (and even reduce) the prediction performance and for this reason, we observe that the gain in terms of prediction performance of NoEC_vs_FE is less pronounced regarding comparing EC_vs_FE.

Discussions

Due to the fact, that still the practical implementation of the GS methodology is challenging since not always is possible to guarantee high genomic-enabled prediction accuracy, many strategies had been developed to improve the machine learning genomic prediction ability (Sallam and Smith, 2016). For this reason, since the GS methodology is still not optimal, this investigation explored FE on the environmental covariates. FE is a crucial step in machine learning and data science that involves creating new features or modifying existing ones to improve the performance of a model. FE is a creative and essential aspect of the machine learning workflow, and it can significantly impact the success of one’s models. It is a skill that improves with experience and a deep understanding of the data and problem. For this reason, FE has been applied successfully in solving natural language processing, computer vision, time series and other issues.

FE is not new in the context of GS, since some studies had been conducted exploring feature engineering techniques from the feature selection point of view. For example, Long et al. (2011) used dimension reduction and variable selection for genomic selection to predict milk yield in Holsteins. Tadist et al. (2019) present a systematic and structured literature review of the feature-selection techniques used in studies related to big genomic data analytics. While Meuwissen et al. (2017) proposed variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. More recently Montesinos-López et al. (2023) proposed feature selection methods for selecting environmental covariables to enhance genomic prediction accuracy. However, these studies are only focused on feature selection and not create new features from the original inputs.

From our results across traits and data sets, we can state that including environmental covariates significantly improves the prediction performance, since comparing no environmental covariates (NoEC) vs adding environmental covariates (EC), the resulting improvement was of 167.900% (RE=2.679 of NoEC_vs_EC), 142.100 (RE=2.242 of NoEC_vs_EC), 56.100% (RE=1.561 of NoEC_vs_EC) and 421.300% (RE=5.213 of NoEC_vs_EC) under predictor E+G, E+G+GE, E+G+BRR and E+G+GE+BRR respectively. However, it is very interesting to point out that the prediction performance can be even improved when the covariates are included but using FE. We found that the improvement of the prediction performance using FE only including only the EC was of 816.600% (RE=9.166 of EC_vs_FE), 372.900% (RE=4.729 of EC_vs_FE), 616.100% (RE=716.100 of EC_vs_FE) and 1240.900% (RE=13.409% of EC_vs_FE) under predictors E+G, E+G+GE, E+G+BRR and E+G+GE+BRR respectively. The larger gain in prediction performance was observed under the most complex predictor (E+G+GE+BRR), while the lowest gain was observed under predictor E+G+GE. Our results show that FE in genomic prediction holds tremendous potential for advancing our understanding of genetics and improving predictions related to various aspects of genomics. For this reason, FE should be considered an important tool to unlock the potential of genomic data for research and practical applications of genomic prediction.

Although our results are very promising for the use of FE, its practical implementation is very challenging, since we observed a significant improvement in some data sets but not in all, and for practical implementations, we need to be able to identify with a high degree of accuracy when the use of FE will be beneficial and when the use of this approach will not be successful. Also, it is important to point out that we have opted against utilizing the Pearson’s correlation coefficient as a performance metric for predicting outcomes. This decision is principally rooted in the lack of substantial improvement linked to this measure we observed. The marginal benefits observed with this metric can be partly ascribed to our exclusive focus on feature selection within the realm of environmental covariates. Additionally, this can be attributed to the assessment of environmental covariates not at the genotype level but rather at the environmental (location) level.

Three reasons why the FE works well for some data but not very well for others are: (1) that those data sets with low efficiency with

FE are those in which the environmental covariates are less correlated with the response variable, (2) that we speculate that not for all data sets the type of FE we implemented are efficient and (3) FE capture complex relationships between the inputs and the response variable. These mean that the nature of each data set affects substantially the performance of any FE strategy. For these reasons some challenges for its implementation are: *a) Domain Knowledge Requirement*: Effective FE often requires a deep understanding of the domain. With domain expertise, it can be easier to identify relevant features or transformations that could enhance model performance; *b) Data Quality and Quantity*: Obtaining high-quality and sufficient data for FE can be challenging in many practical scenarios. Limited or noisy data can hinder the creation of meaningful features; *c) Time and Resource Constraints*: Implementing FE can be time-consuming, and in some real-world applications, there might be strict time and resource constraints. This makes exploring and experimenting with a wide range of FE techniques challenging; *d) Dynamic Data*: Real-world data often changes over time. Features that are effective at one point in time may become less relevant or even obsolete as the data distribution evolves. Maintaining and updating features in dynamic environments can be challenging; *e) Overfitting Risks*: Aggressive FE can lead to overfitting, especially when the number of features is large compared to the amount of available data. Overfit models perform well on training data but generalize poorly to new, unseen data; *f) Complexity and Interpretability*: As the number and complexity of features increase, the resulting models can become difficult to interpret. This lack of interpretability can be challenging, especially in applications where understanding the model's decisions is crucial; *g) Automated Feature Selection*: While manual FE can be effective, the process is often subjective and time-consuming. Automated feature selection methods exist, but selecting the right techniques and parameters can be challenging; *h) Curse of Dimensionality*: As the number of features increases, the curse of dimensionality becomes more pronounced. This can lead to increased computational requirements and decreased model performance, making it challenging to strike the right balance.

The results of this study demonstrate that the feature engineering strategy for incorporating environmental covariates effectively enhances genomic prediction accuracy. However, further research is warranted to refine the methodology for integrating environmental covariates into genomic prediction models, particularly in the context of modeling genotype-environment interactions (GE). For instance, employing the factor analytic (FA) multiplicative operator to describe cultivar effects in different environments has shown promise as a robust and efficient machine learning approach for analyzing multi-environment breeding trials (Piepho, 1998; Smith et al., 2005). Factor analysis offers solutions for modeling GE with heterogeneous variances and covariances, either alongside the numerical relationship matrix (based on pedigree information) (Crossa et al., 2006) or utilizing the genomic similarity matrix to assess GE (Burgueño et al., 2012). Further research is needed to comprehensively explore the application of the FA approach for feature engineering of environmental covariates within the framework of genomic prediction.

Conclusions

This study delved into the impact of feature engineering on environmental covariates to enhance the predictive capabilities of genomic models. Our findings demonstrate a consistent improvement in prediction performance, as measured by MSE, across most datasets when employing feature engineering techniques compared to models without such enhancements. While some datasets showed no significant gains, others exhibited notably substantial improvements. These results underscore the potential of feature engineering to bolster prediction accuracy in genomic studies. However, it's imperative to acknowledge the inherent complexity and challenges associated with practical implementation, as various factors can influence its efficacy. Therefore, we advocate for further exploration and adoption of feature engineering methodologies within the scientific community to accumulate more empirical evidence and harness its full potential in genomic prediction.

Data availability statement

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

Author contributions

OM: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. LC: Writing – review & editing, Conceptualization. CS: Writing – review & editing, Supervision, Project administration, Investigation. BC: Writing – review & editing, Software, Methodology, Formal Analysis, Data curation, Conceptualization. GH: Writing – review & editing, Software, Conceptualization. BA: Writing – review & editing, Software, Methodology, Investigation, Data curation. SR: Writing – review & editing, Software, Methodology, Investigation. GG: Writing – review & editing, Methodology, Investigation, Data curation. KA: Writing – review & editing, Methodology, Investigation. RF: Writing – review & editing, Methodology, Investigation, Conceptualization. AM: Writing – review & editing, Software, Methodology, Investigation, Conceptualization. JC: Writing – review & editing, Writing – original draft, Investigation, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Open Access fees were received from the Bill & Melinda Gates Foundation. We acknowledge the financial support provided by the Bill & Melinda Gates Foundation (INV-003439 BMGF/FCDO Accelerating Genetic Gains in Maize and Wheat for Improved Livelihoods (AGG)) as well as the USAID projects (Amend. No. 9 MTO 069033, USAID-CIMMYT Wheat/AGGMW, Genes 2023, 14, 927 14 of 18AGG-Maize Supplementary Project, AGG (Stress Tolerant Maize for Africa))

which generated the CIMMYT data analyzed in this study. We are also thankful for the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) through the Research Council of Norway for grants 301835 (Sustainable Management of Rust Diseases in Wheat) and 320090 (Phenotyping for Healthier and more Productive Wheat Crops). We acknowledge the support of the Window 1 and 2 funders to the Accelerated Breeding Initiative (ABI).

Conflict of interest

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

The handling editor HL declared a past co-authorship with the author JC.

References

- Abed, A., Pérez-Rodríguez, P., Crossa, J., and Belzile, F. (2018). When less can be better: how can we make genomic selection more cost-effective and accurate in Barley? *Theor. Appl. Genet.* 131, 1873–1890. doi: 10.1007/s00122-018-3120-8
- Afshar, M., and Usefi, H. (2020). High-dimensional feature selection for genomic datasets. *Knowledge-Based Syst.* 206, 106370. doi: 10.1016/j.knsys.2020.106370
- Akdemir, D., Sanchez, J. I., and Jannink, J. L. (2015). Optimization of genomic selection training populations with a genetic algorithm. *Genet. Selection Evolution.* 47, 1–10. doi: 10.1186/s12711-015-0116-6
- Bassi, F. M., Bentley, A. R., Charmet, G., Ortiz, R., and Crossa, J. (2016). Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). *Plant Sci.* 242, 23–36. doi: 10.1016/j.plantsci.2015.08.021
- Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci. Rep.* 5, 10312. doi: 10.1038/srep10312
- Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Sci.* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690
- Beyene, Y., Semagn, K., Mugo, S., Tarekegne, A., Babu, R., Meisel, B., et al. (2015). Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. *Crop Sci.* 55, 154–163. doi: 10.2135/cropsci2014.07.0460
- Budhlakoti, N., Kushwaha, A. K., Rai, A., Chaturvedi, K. K., Kumar, A., Pradhan, A. K., et al. (2022). Genomic selection: A tool for accelerating the efficiency of molecular breeding for development of climate-resilient crops. *Front. Genet.* 13. doi: 10.3389/fgene.2022.832153
- Burgueño, J., de los Campos, G., Weigel, K., and José Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype \times environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719. doi: 10.2135/cropsci2011.06.0299
- Butoto, E. N., Brewer, J. C., and Holland, J. B. (2022). Empirical comparison of genomic and phenotypic selection for resistance to Fusarium ear rot and fumonisin contamination in maize. *Theor. Appl. Genet.* 135, 2799–2816. doi: 10.1007/s00122-022-04150-8
- Calus, M. P. L., and Veerkamp, R. F. (2011). Accuracy of multi-trait genomic selection using different methods. *Genet. Sel. Evol.* 43, 1–14. doi: 10.1186/1297-9686-43-26
- Carrillo-de-Albornoz, J., Rodríguez Vidal, J., and Plaza, L. (2018). Feature engineering for sentiment analysis in e-health forums. *PloS One* 13, e0207996. doi: 10.1371/journal.pone.0207996
- Costa-Neto, G., Crossa, J., and Fritsche-Neto, R. (2021b). Enviromic assembly increases accuracy and reduces costs of the genomic prediction for yield plasticity in maize. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.717552
- Costa-Neto, G., Fritsche-Neto, R., and Crossa, J. (2021a). Nonlinear kernels, dominance, and envirotyping data increase the accuracy of genome-based prediction in multi-environment trials. *Heredity* 126, 92–106. doi: 10.1038/s41437-020-00353-1
- Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., and Krishnamachari, A. (2006). Modeling genotype \times environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Sci.* 46, 1722–1733. doi: 10.2135/cropsci2005.11-0427
- Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011
- Desta, Z. A., and Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci.* 19, 592–601. doi: 10.1016/j.tplants.2014.05.006
- Dong, G., and Liu, H. (2018). *Feature engineering for machine learning and data analytics* (California, USA: CRC press).
- FAO. (2023). “The state of food security and nutrition in the world 2023,” in *Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum* (Rome, Italy: FAOSTAT).
- Gesteiro, N., Ordás, B., Butrón, A., de la Fuente, M., Jiménez-Galindo, J. C., Samayoa, L. F., et al. (2023). Genomic versus phenotypic selection to improve corn borer resistance and grain yield in maize. *Front. Plant Sci.* 14. doi: 10.3389/fpls.2023.1162440
- Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190
- Heaton, J. (2016). “An empirical analysis of feature engineering for predictive modeling,” in *SoutheastCon 2016*. (pp. 1–(pp.6) (Norfolk, Virginia, USA: IEEE). doi: 10.1109/SECON.2016.7506650
- Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Sci.* 49, 1–12. doi: 10.2135/cropsci2008.08.0512
- Hu, H., Campbell, M. T., Yeats, T. H., Zheng, X., Runcie, D. E., Covarrubias-Pazarán, G., et al. (2021). Multi-omics prediction of oat agronomic and seed nutritional traits across environments and in distantly related populations. *Theor. Appl. Genet.* 134 (12), 4043–4054. doi: 10.1007/s00122-021-03946-4
- Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theor. Appl. Genet.* 127 (3), 595–607. doi: 10.1007/s00122-013-2243-1
- Jarquín, D., de Leon, N., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I., et al. (2020). Utility of climatic information via combining ability models to improve genomic prediction for yield within the genomes to fields maize project. *Front. Genet.* 11, 592769. doi: 10.3389/fgene.2020.592769
- Juliana, P., Singh, R. P., Poland, J., Mondal, S., Crossa, J., Montesinos-López, O. A., et al. (2018). Prospects and challenges of applied genomic selection—A new paradigm in breeding for grain yield in bread wheat. *Plant Genome* 11, 1–17. doi: 10.3835/plantgenome2018.03.0017
- Khurana, U., Samulowitz, H., and Turaga, D. (2018). “Feature engineering for predictive modeling using reinforcement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence* (New Orleans, LA, USA), Vol. 32.
- Krause, M. R., González-Pérez, L., Crossa, J., Pérez-Rodríguez, P., Montesinos-López, O., Singh, R. P., et al. (2019). Hyperspectral reflectance derived relationship matrices for genomic prediction of grain yield in wheat. *G3 Genes Genomes Genet.* 9, 1231–1247. doi: 10.1534/g3.118.200856
- Lam, H. T., Thiebaut, J. M., Sinn, M., Chen, B., Mai, T., and Alkan, O. (2017). One button machine for automating feature engineering in relational databases. *arXiv*. doi: 10.48550/arXiv.1706.00327
- Lawrence-Dill, C. J. (2017) *Genomes to fields: GxE Field Experiment*. Available online at: <https://www.genomes2fields.org/resources/#sop> (Accessed 2021 January 11).
- Long, N., Gianola, D., Rosa, G. J., and Weigel, K. A. (2011). Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *J. Anim. Breed Genet.* 128, 247–257. doi: 10.1111/jbg.2011.128.issue-4

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.2024.1349569/full#supplementary-material>

- Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157 (12), 1819–1829. doi: 10.1093/GENETICS/157.4.1819
- Meuwissen, T. H. E., Indahl, U. G., and Ødegård, J. (2017). Variable selection models for genomic selection using whole-genome sequence data and singular value decomposition. *Genet. Sel. Evol.* 49, 94. doi: 10.1186/s12711-017-0369-3
- Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W. R., Fajardo-Flores, S. B., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22 (1), 19. doi: 10.1186/s12864-020-07319-x
- Montesinos-López, O. A., Montesinos-López, A., and Crossa, J. (2022). *Multivariate statistical Machine Learning Methods for Genomic Prediction* (Cham, Switzerland: Springer International Publishing). doi: 10.1007/978-3-030-89010-0
- Montesinos-López, O. A., Crespo-Herrera, A., Saint Pierre, J., Bentley, J., de la Rosa-Santamaria, J., Ascencio-Laguna, J., et al. (2023). Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy? *Front. Genet.* 14, 1209275. doi: 10.3389/fgene.2023.1209275
- Montesinos-López, A., Montesinos-López, O. A., Cuevas, J., Mata-López, W. A., Burgueño, J., Mondal, S., et al. (2017). Genomic Bayesian functional regression models with interactions for predicting wheat grain yield using hyper-spectral image data. *Plant Methods* 13, 1–29. doi: 10.1186/s13007-017-0212-4
- Monteverde, E., Gutierrez, L., Blanco, P., Pérez de Vida, F., Rosas, J. E., Bonnacarrère, V., et al. (2019). Integrating molecular markers and environmental covariates to interpret genotype by environment interaction in rice (*Oryza sativa* L.) grown in subtropical areas. *G3 (Bethesda)* 9, 1519–1531. doi: 10.1534/g3.119.400064
- Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., and Turaga, D. S. (2017). “Learning feature engineering for classification,” in *Ijcai* (Melbourne Australia), Vol. 17. 2529–2535.
- Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442
- Piepho, H. P. (1998). Empirical best linear unbiased prediction in cultivar trials using factor analytic variance covariance structure. *Theor. Appl. Genet.* 97, 195–201. doi: 10.1007/s001220050885
- Rincint, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., et al. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics* 192, 715–728. doi: 10.1534/genetics.112.141473
- Rogers, A. R., Dunne, J. C., Romay, C., Bohn, M., Buckler, E. S., Ciampitti, I. A., et al. (2021). The importance of dominance and genotype-by-environment interactions on grain yield variation in a large-scale public cooperative maize experiment. *G3 (Bethesda)* 11, jkaa050. doi: 10.1093/g3journal/jkaa050
- Rogers, A. R., and Holland, J. B. (2022). Environment-specific genomic prediction ability in maize using environmental covariates depends on environmental similarity to training data. *G3 Genes|Genomes|Genetics* 12, jkab440. doi: 10.1093/g3journal/jkab440
- Sallam, A. H., and Smith, K. P. (2016). Genomic selection performs similarly to phenotypic selection in Barley. *Crop Sci.* 56, 2871–2881. doi: 10.2135/cropsci2015.09.0557
- Smith, A. B., Cullis, B. R., and Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: An overview of current mixed model approaches. *J. Agric. Sci.* 143, 1–14. doi: 10.1017/S0021859605005587
- Tadist, K., Najah, S., Nikolov, N. S., Mrabti, F., and Zahi, A. (2019). Feature selection methods and genomic big data: a systematic review. *J. Big Data* 6, 79. doi: 10.1186/s40537-019-0241-0
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980
- Wang, K., Abid, M. A., Rasheed, A., et al. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Mol. Plant* 16, 279–293. doi: 10.1016/j.molp.2022.11.004
- Wu, P. Y., Stich, B., Weisweiler, M., Shrestha, A., Erban, A., Westhoff, P., et al. (2022). Improvement of prediction ability by integrating multi-omic datasets in barley. *BMC Genomics* 23, 200. doi: 10.1186/s12864-022-08337-7
- Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., et al. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Commun.* 1, 100005. doi: 10.1016/j.xplc.2019.100005
- Yurek, O. E., and Birant, D. (2019). “Remaining useful life estimation for predictive maintenance using feature engineering,” in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. 1–5 (Izmir, Turkeyand: IEEE). doi: 10.1109/ASYU48272.2019.894639

Appendix A

Bayesian ridge regression

Bayesian Ridge Regression (BRR) is a probabilistic approach to linear regression that incorporates Bayesian principles. It is a regularized regression method that extends traditional linear regression by introducing a prior distribution over the regression coefficients. This approach provides a way to express uncertainty in the model parameters and helps prevent overfitting by introducing regularization.

The model assumptions assumes a traditional linear regression, with a linear relationship between the independent variables and the dependent variable. The BRR assumes that the coefficients of the regression model follow a Gaussian (normal) distribution. This introduces a regularization term that penalizes large coefficients, helping to prevent overfitting.

The model formulation assumes that X is an independent variables with and a dependent variable y , such that the BRR can be written as

$$y = X\beta + \epsilon$$

where y is the dependent variable. X is the matrix of independent variables, β is the vector of regression coefficients and ϵ is the residual (error) term. From a Bayesian perspective, the prior distribution for β is assumed to be Gaussian (normal) $\beta \sim N(0, \alpha^{-1}I)$ with α being a hyperparameter controlling the strength of the regularization and I is the identity matrix. The goal is to estimate the posterior distribution of β given the data. The posterior distribution is proportional to the product of the likelihood and the prior $P(\beta | X, y) \propto P(y | X, \beta) \cdot P(\beta)$. Once the posterior distribution is obtained, Bayesian inference can be performed with point estimates (mean or mode) of the posterior distribution can be used as the regression coefficients. additionally, credible intervals can be computed to quantify uncertainty.

Appendix B

Japonica dataset

Predictor: E+G

Table B1 shows an adequate performance for the results under NoEC for the GC trait across all environments. The MSE values for 2009, 2010, 2011, 2012, and 2013 were 0.0035, 0.0110, 0.0019, 0.0281, and 0.0017, respectively. Comparing the NoEC results to the EC and FE techniques using Relative Efficiency (RE), all RE values were below 1. On average, NoEC presented 50.050% better performance compared to EC and 42.230% better performance compared to FE. However, when comparing EC and FE techniques based on RE, FE outperformed EC in 2010, 2011, 2012, and 2013, with RE values of 1.287, 2.686, 1.139, and 1.586, respectively. In 2009, EC had a lower RE value of 0.522. On average, the use of FE outperformed EC by 44.410%. Please refer to Table B1 for more detailed information.

Concerning the GY trait, Table B1 shows that the use of EC led to a superior performance in most environments based on MSE (796,963 [2009], 2,488,872 [2010] and 1,157,280 [2012]). However, the exceptions occurred in 2011 and 2013, when FE achieved the best MSE values of 2,615,758 and 377,719, respectively. By contrast, when comparing NoEC versus EC and NoEC versus FE using RE, most RE values were greater than 1. On average, the EC technique displayed an improvement of 105.610% (NoEC_vs_EC) regarding the NoEC method, and an improvement of 77.570% (NoEC_vs_FE) was observed with the use of FE compared to the conventional NoEC technique. Nonetheless, when assessing the performance of EC and FE techniques based on RE, FE only outperformed EC in 2011 (RE = 1.091) and 2013 (RE = 1.087). EC, on the other hand, outperformed FE in 2009 (RE = 0.777), 2010 (RE = 0.817), and 2012 (RE = 0.806), resulting in an average RE of 0.916. This indicates an overall performance loss of 8.450% when using FE compared to EC. Table B1 provides further details.

In terms MSE for the PH trait, Table B1 shows that the use of FE achieved the best performance in most environments (15.872 [2009], 10.959 [2010], and 164.039 [2012]). However, there were exceptions in 2011 and 2013, where the best MSE values were 28.573 (EC) and 18.363 (NoEC), respectively. On the other hand, when comparing NoEC versus EC and NoEC versus FE techniques using RE, most RE values were greater than 1. On average, the use of EC and FE displayed improvements of 61.570% and 70.210%, respectively, compared to the use of NoEC. Furthermore, when comparing the performance of EC and FE techniques based on RE, FE outperformed EC in all environments, resulting in an average RE of 1.0389. This indicates that using FE surpassed EC by 3.88% (Table B1).

In terms of MSE for the PHR trait, Table B1 indicates that the use of FE yielded the best performance in most environments (0.001 [2009], 0.001 [2010], and 0.001[2013]). However, exceptions were found in 2011 and 2012, when the best MSE values were 0.001 (EC) and 0.006 (NoEC), respectively. On the other hand, when comparing EC versus FE and NoEC versus FE techniques using Relative Efficiency (RE), most RE values were at least 1. On average, the use of FE displayed a general improvement of 22.790%, compared to EC and 7.020% compared to the conventional NoEC technique. However, evaluating the performance of EC versus NoEC techniques based on RE showed that NoEC outperformed EC in most environments, resulting in an average RE of 0.938. This indicates a general accuracy loss of 6.200% when using EC compared to the conventional NoEC technique (Table B1).

Predictor: E+G+GE

Table B1 shows that, in most environments, the conventional NoEC technique yielded the best performance for the GC trait, with MSE values of 0.001 (2009), 0.013 (2010), and 0.002 (2011). The exceptions occurred in 2012 and 2013, with the best MSE values of 0.025 (EC) and 0.0023 (FE). The average RE for the comparison of NoEC versus EC and NoEC versus FE techniques across environments was 0.919 and 0.9023, respectively, indicating general losses of 8.080% and 9.740% for EC and FE compared to the conventional NoEC.

TABLE B1 The prediction performance and the relative efficiency (RE) for Japonica dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictors E+G and E+G+GE under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates with Feature Engineering (FE).

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G	GC	2009	0.004	0.005	0.009	0.729	0.522	0.380
E+G	GC	2010	0.011	0.017	0.013	0.663	1.287	0.853
E+G	GC	2011	0.002	0.009	0.004	0.202	2.686	0.543
E+G	GC	2012	0.028	0.039	0.034	0.719	1.140	0.819
E+G	GC	2013	0.002	0.009	0.006	0.185	1.586	0.293
E+G	GC	Across	–	–	–	0.500	1.444	0.578
E+G	GY	2009	3049246.325	796963.009	1025847.337	3.826	0.777	2.972
E+G	GY	2010	5683515.755	2488872.780	3046722.045	2.284	0.817	1.866
E+G	GY	2011	4024422.454	2853854.731	2615758.363	1.410	1.091	1.539
E+G	GY	2012	2050745.031	1157280.313	1436429.272	1.772	0.806	1.428
E+G	GY	2013	405886.860	410565.496	377719.356	0.989	1.087	1.075
E+G	GY	Across	–	–	–	2.056	0.916	1.776
E+G	PH	2009	58.674	16.561	15.872	3.543	1.043	3.697
E+G	PH	2010	27.005	12.127	10.959	2.227	1.107	2.464
E+G	PH	2011	13.534	28.641	28.573	0.473	1.002	0.474
E+G	PH	2012	175.254	168.840	164.039	1.038	1.029	1.068
E+G	PH	2013	18.363	23.009	22.729	0.798	1.012	0.808
E+G	PH	Across	–	–	–	1.616	1.039	1.702
E+G	PHR	2009	0.001	0.001	0.001	0.750	1.333	1.000
E+G	PHR	2010	0.001	0.002	0.001	0.750	1.600	1.200
E+G	PHR	2011	0.002	0.001	0.002	1.643	0.778	1.278
E+G	PHR	2012	0.006	0.007	0.006	0.797	1.095	0.873
E+G	PHR	2013	0.001	0.001	0.001	0.750	1.333	1.000
E+G	PHR	Across	–	–	–	0.938	1.228	1.070
E+G+GE	GC	2009	0.001	0.001	0.003	0.769	0.433	0.333
E+G+GE	GC	2010	0.013	0.034	0.032	0.394	1.053	0.414
E+G+GE	GC	2011	0.002	0.006	0.003	0.281	2.462	0.692
E+G+GE	GC	2012	0.025	0.025	0.029	1.004	0.839	0.843
E+G+GE	GC	2013	0.006	0.003	0.003	2.148	1.039	2.231
E+G+GE	GC	Across	–	–	–	0.919	1.165	0.903
E+G+GE	GY	2009	3242702.030	1152261.036	1460144.165	2.814	0.789	2.221
E+G+GE	GY	2010	4339466.437	3653811.519	4302236.223	1.188	0.849	1.009
E+G+GE	GY	2011	1834248.259	3337540.514	3251492.136	0.550	1.027	0.564
E+G+GE	GY	2012	1894112.619	989127.176	1358843.398	1.915	0.728	1.394
E+G+GE	GY	2013	1924915.862	416054.225	370980.321	4.627	1.122	5.189
E+G+GE	GY	Across	–	–	–	2.219	0.903	2.075
E+G+GE	PH	2009	56.517	20.261	17.631	2.789	1.149	3.206

(Continued)

TABLE B1 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE	PH	2010	17.957	12.954	16.142	1.386	0.803	1.112
E+G+GE	PH	2011	44.689	77.310	64.564	0.578	1.197	0.692
E+G+GE	PH	2012	164.891	175.005	168.680	0.942	1.038	0.978
E+G+GE	PH	2013	59.136	24.696	23.544	2.395	1.049	2.512
E+G+GE	PH	Across	–	–	–	1.618	1.047	1.700
E+G+GE	PHR	2009	0.001	0.001	0.001	0.750	1.333	1.000
E+G+GE	PHR	2010	0.002	0.002	0.002	0.818	1.467	1.200
E+G+GE	PHR	2011	0.002	0.001	0.001	1.727	0.917	1.583
E+G+GE	PHR	2012	0.005	0.007	0.006	0.783	1.095	0.857
E+G+GE	PHR	2013	0.001	0.001	0.001	0.750	1.600	1.200
E+G+GE	PHR	Across	–	–	–	0.966	1.282	1.168

Regarding the GY trait, MSE values from Table B1 reveal that the use of EC achieved the best performance in most environments (1152261.030 [2009], 3653811.510 [2010], and 989127.170 [2012]). However, exceptions were observed in 2011 and 2013, where the best MSE values were 1834248.25 (NoEC) and 30980.32 (FE), respectively. On the other hand, when comparing NoEC versus EC and NoEC versus FE techniques using RE, most RE values were greater than 1. The average RE for NoEC versus EC and NoEC versus FE was 2.219 and 2.075, respectively, indicating general improvements of 121.860% and 107.520% compared to the use of NoEC. However, an evaluation of the performance of EC and FE techniques based on RE showed that FE outperformed EC only in 2011 (1.0267) and 2013 (1.122), while EC outperformed FE in 2009 (0.789), 2010 (0.849), and 2012 (0.7278). Consequently, the average RE for EC versus FE was 0.9029, implying a general loss of 9.710% when using FE compared to EC (Table B1).

Concerning the PH trait, the analysis of MSE values from Table B1 reveals that the use of FE yielded the best performance in most environments (17.631 [2009] and 23.544 [2012]). However, exceptions were observed in 2010, 2011, and 2013, where the best MSE values were 12.954 (EC), 44.689 (NoEC), and 164.891 (NoEC), respectively. On the other hand, comparing NoEC versus EC and NoEC versus FE techniques using RE showed that most RE values were greater than 1. The average RE for NoEC versus EC and NoEC versus FE was 1.618 and 1.700, respectively, indicating general improvements of 61.810% and 70.000% compared to the conventional NoEC technique. Furthermore, when evaluating the performance of EC and FE techniques based on RE, FE consistently outperformed EC in most environments. The average RE for EC versus FE was 1.047, indicating a 4.710% advantage in favor of FE (Table B1).

Moreover, in the case of the PHR trait, the analysis of MSE values from Table B1 shows that the use of FE yielded the best performance in most environments (0.001 [2009], 0.002 [2010], and 0.001 [2013]). However, there were exceptions in 2011 and 2012, where the best MSE values were 0.001 (EC) and 0.005 (NoEC), respectively. Furthermore, when comparing the RE values between NoEC versus EC and NoEC versus FE techniques, the average RE values of 0.966 and 1.168 indicate

a slight loss of 3.440% and an improvement of 16.800%, respectively, for the use of EC and FE compared to the conventional NoEC technique. Nevertheless, when evaluating the performance of FE versus EC techniques based on RE, FE consistently outperformed EC in most environments. The average RE for FE versus EC was 1.282, indicating a significant improvement of 28.240% in accuracy for using FE compared to (Table B1).

Predictor: E+G+BRR

According to Table B2, the GC trait displayed superior performances with the conventional NoEC technique in most environments, yielding MSE values of 0.004 (2009), 0.002 (2011), and 0.0012 (2013). However, exceptions were found in 2010 and 2012, where FE achieved the best MSE values of 0.0680 and 0.009, respectively. Comparing the RE values between NoEC versus EC and NoEC versus FE techniques showed that most RE values were below 1. Nonetheless, the average RE of 1.104 (NoEC_vs_EC) and 1.189 (NoEC_vs_FE) indicated that EC and FE outperformed the conventional NoEC technique by 10.360% and 18.930%, respectively. Furthermore, when evaluating the performance of EC and FE techniques based on RE, FE presented the best performance in 2009 (1.151), 2010 (1.353), 2011 (2.044), and 2012 (1.0623), while EC outperformed FE in 2013 (0.529). Overall, the average RE 1.228 indicated that FE outperformed EC by 22.800% (Table B2).

Regarding the GY trait, Table B2 indicates that the conventional NoEC technique displayed superior performances in most environments, with MSE values of 5,683,515.750 (2010), 2,749,626.080 (2012), and 405,886.860 (2013). However, exceptions were observed in 2009 and 2011, where FE achieved the best MSE values of 3,049,246.320 and 4,024,422.450, respectively. When comparing the RE values between NoEC_vs_EC and NoEC_vs_FE techniques, most values were below 1. Nevertheless, the average RE of 1.124 (NoEC_vs_EC) and 0.896 (NoEC_vs_FE) indicated an overall improvement of 12.430% for EC and a general loss of 10.450% for FE compared to the conventional NoEC technique. However, when comparing the performance of EC and FE techniques based on RE,

TABLE B2 The prediction performance and the relative efficiency (RE) for Japonica dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictors E+G+BRR and E+G+GE+BRR under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates *with Feature Engineering (FE)*.

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+BRR	GC	2009	0.004	0.008	0.007	0.417	1.151	0.480
E+G+BRR	GC	2010	0.011	0.009	0.007	1.196	1.353	1.618
E+G+BRR	GC	2011	0.002	0.009	0.005	0.207	2.044	0.422
E+G+BRR	GC	2012	0.028	0.010	0.010	2.755	1.063	2.927
E+G+BRR	GC	2013	0.002	0.002	0.003	0.944	0.529	0.500
E+G+BRR	GC	Across	–	–	–	1.104	1.228	1.189
E+G+BRR	GY	2009	3049246.325	1221342.669	1607864.482	2.497	0.760	1.897
E+G+BRR	GY	2010	5683515.755	7662804.296	7449307.222	0.742	1.029	0.763
E+G+BRR	GY	2011	4024422.454	3689043.326	3841776.983	1.091	0.960	1.048
E+G+BRR	GY	2012	2050745.031	2749626.084	5697594.878	0.746	0.483	0.360
E+G+BRR	GY	2013	405886.860	743092.012	988735.462	0.546	0.752	0.411
E+G+BRR	GY	Across	–	–	–	1.124	0.797	0.896
E+G+BRR	PH	2009	58.674	15.466	15.281	3.794	1.012	3.840
E+G+BRR	PH	2010	27.005	22.962	27.436	1.176	0.837	0.984
E+G+BRR	PH	2011	13.534	29.033	25.921	0.466	1.120	0.522
E+G+BRR	PH	2012	175.254	165.479	159.312	1.059	1.039	1.100
E+G+BRR	PH	2013	18.363	10.981	14.450	1.672	0.760	1.271
E+G+BRR	PH	Across	–	–	–	1.634	0.954	1.543
E+G+BRR	PHR	2009	0.001	0.001	0.001	1.000	1.000	1.000
E+G+BRR	PHR	2010	0.001	0.001	0.001	1.714	1.167	2.000
E+G+BRR	PHR	2011	0.002	0.001	0.001	2.875	0.889	2.556
E+G+BRR	PHR	2012	0.006	0.006	0.011	0.887	0.554	0.491
E+G+BRR	PHR	2013	0.001	0.001	0.001	1.200	1.000	1.200
E+G+BRR	PHR	Across	–	–	–	1.535	0.922	1.449
E+G+GE+BRR	GC	2009	0.001	0.007	0.006	0.154	1.083	0.167
E+G+GE+BRR	GC	2010	0.013	0.017	0.008	0.796	2.012	1.602
E+G+GE+BRR	GC	2011	0.002	0.007	0.003	0.273	2.000	0.546
E+G+GE+BRR	GC	2012	0.025	0.024	0.019	1.029	1.278	1.316
E+G+GE+BRR	GC	2013	0.006	0.004	0.002	1.526	2.111	3.222
E+G+GE+BRR	GC	Across	–	–	–	0.756	1.697	1.371
E+G+GE+BRR	GY	2009	3242702.030	1333530.864	1860560.276	2.432	0.717	1.743
E+G+GE+BRR	GY	2010	4339466.437	7649947.049	7468881.672	0.567	1.024	0.581
E+G+GE+BRR	GY	2011	1834248.259	4157537.398	4872981.083	0.441	0.853	0.376
E+G+GE+BRR	GY	2012	1894112.619	1690390.524	4082192.704	1.121	0.414	0.464
E+G+GE+BRR	GY	2013	1924915.862	584945.854	681359.148	3.291	0.859	2.825
E+G+GE+BRR	GY	Across	–	–	–	1.570	0.773	1.198
E+G+GE+BRR	PH	2009	56.517	18.089	17.332	3.124	1.044	3.261

(Continued)

TABLE B2 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE+BRR	PH	2010	17.957	14.956	26.970	1.201	0.555	0.666
E+G+GE+BRR	PH	2011	44.689	22.351	22.026	1.999	1.015	2.029
E+G+GE+BRR	PH	2012	164.891	171.095	167.745	0.964	1.020	0.983
E+G+GE+BRR	PH	2013	59.136	11.071	12.138	5.342	0.912	4.872
E+G+GE+BRR	PH	Across	–	–	–	2.526	0.909	2.362
E+G+GE+BRR	PHR	2009	0.001	0.001	0.001	1.000	1.000	1.000
E+G+GE+BRR	PHR	2010	0.002	0.001	0.001	2.571	1.000	2.571
E+G+GE+BRR	PHR	2011	0.002	0.001	0.001	2.375	1.000	2.375
E+G+GE+BRR	PHR	2012	0.005	0.006	0.011	0.871	0.554	0.482
E+G+GE+BRR	PHR	2013	0.001	0.001	0.001	1.200	0.833	1.000
E+G+GE+BRR	PHR	Across	–	–	–	1.604	0.877	1.486

only FE presented a superior performance in 2010 (1.029), resulting in an average RE of 0.797, which indicates a general loss of 20.350% for FE compared to EC (Table B2).

For the PH trait, Table B2 shows that FE yielded the best performance in environments 2009 (15.281) and 2012 (159.312), while EC led to superior performances in environments 2010 (22.962) and 2013 (10.981). Most notably, when comparing the RE values for NoEC_vs_EC and NoEC_vs_FE, values exceeding 1 were observed. The average RE values of 1.634 (NoEC_vs_EC) and 1.5434 (NoEC_vs_FE) indicated substantial improvements of 63.350% and 54.350% respectively for using EC and FE, compared to the conventional NoEC technique. However, in evaluating the performance of EC and FE based on RE, FE exhibited a superior performance in most environments, but still resulting in an average RE of 0.954. This suggests that EC marginally outperformed FE by 4.650%. For further details, see Table B2.

Additionally, for the PHR trait, using FE displayed a superior performance in most environments, as indicated in Table B2. The best MSE values were observed in 2009 (0.001), 2010 (0.001), and 2013 (0.001). However, exceptions were noted in 2011 and 2012, where the use of EC and NoEC resulted in the best MSE values of 8×10^{-4} and 0.0055, respectively. Furthermore, most RE values comparing NoEC_vs_EC and NoEC_vs_FE techniques were greater than 1. The average RE values of 1.535 (NoEC_vs_EC) and 1.449 (NoEC_vs_FE) indicate significant improvements of 53.530% and 44.930% respectively, compared to the conventional NoEC technique. However, when comparing the performance of the EC versus the FE techniques, the RE values were lower than 1 in most environments, resulting in an average RE of 0.9212. This suggests a general accuracy loss of 7.820% in for using FE compared to using the EC technique (Table B2).

Predictor: E+G+GE+BRR

According to Table B2, the GC trait displayed superior performances with the conventional NoEC technique in most environments, yielding MSE values of 0.004 (2009), 0.002 (2011), and 0.0012 (2013). However, exceptions were found in 2010 and

2012, where FE achieved the best MSE values of 0.0680 and 0.009, respectively. Comparing the RE values between NoEC versus EC and NoEC versus FE techniques showed that most RE values were below 1. Nonetheless, the average RE of 1.104 (NoEC_vs_EC) and 1.189 (NoEC_vs_FE) indicated that EC and FE outperformed the conventional NoEC technique by 10.360% and 18.930%, respectively. Furthermore, when evaluating the performance of EC and FE techniques based on RE, FE presented the best performance in 2009 (1.151), 2010 (1.353), 2011 (2.044), and 2012 (1.0623), while EC outperformed FE in 2013 (0.529). Overall, the average RE 1.228 indicated that FE outperformed EC by 22.800% (Table B2).

Regarding the GY trait, the analysis in Table B2 reveals that the use of EC yielded superior results in most environments (2009 [1333530.864], 2012 [1690390.524], and 2013 [584945.854]). However, exceptions were observed in 2010 and 2011, where the NoEC approach resulted in the best MSE values of 4339466.437 and 1834248.259, respectively. Moreover, most RE values for the comparison of NoEC_vs_EC and NoEC_vs_FE techniques were greater than 1. The average RE values of 1.570 (NoEC_vs_EC) and 1.198 (NoEC_vs_FE) indicate general improvements of 57.030% and 19.790% for the use of EC and FE, respectively, compared to the use of NoEC. However, when comparing the performance of EC and FE techniques based on RE, the FE technique did not outperform EC only in 2010, resulting in an average RE of 0.773. This suggests a general loss of 22.670% accuracy for using FE compared to EC.

Regarding the PH trait, Table B2 shows that the use of FE achieved the best performance in environments 2009 (17.332) and 2011 (22.026), while the use of EC achieved the best performance in environments 2010 (14.9561) and 2013 (11.071). Similarly, most of the RE values for the comparison of NoEC_vs_EC and NoEC_vs_FE techniques were greater than 1. The average RE values of 2.5259 (NoEC_vs_EC) and 2.362 (NoEC_vs_FE) indicate general improvements of 152.590% and 136.210% for using EC and FE, respectively, compared to the conventional NoEC technique. However, when comparing the performance of

EC and FE techniques based on RE, EC outperformed FE in most environments, resulting in an average RE of 0.909. This indicates that using EC achieved a 9.100% improvement compared to using FE. For more detailed information, refer to Table 2.

Table B2 displays that using EC yielded the best performance for the PHR trait in most environments, as indicated by the MSE. Specifically, the MSE values were as follows: 2009 (0.001), 2010 (0.001), 2011 (0.001), and 2013 (0.001). However, in 2012, the best MSE values were 0.005, achieved using both EC and NoEC. Comparing NoEC_vs_EC and NoEC_vs_FE techniques, most RE values were at least 1, with average improvements of 60.350% and 48.570% when using EC and FE, respectively, compared to NoEC. Conversely, when comparing EC versus FE techniques, most environments resulted in an average RE of 0.877, indicating a 12.260% decrease in accuracy when using FE compared to EC (Table B2).

USP dataset

Predictor: E+G

Upon examining Table B3, it becomes apparent that the conventional NoEC technique achieved the best performance in terms of MSE in environments Env2 (4.073) and Env3 (5.246). However, exceptions were found in Env1 and Env4, where the optimal MSE values were 3.141 (FE) and 7.814 (EC), respectively. For further detail, refer to Table B3.

Table B3 present our comparison results between the NoEC and EC techniques, assessed through the RE metric. The EC technique displayed its best performance in environments Env1 (1.059) and Env4 (1.046), showcasing improvements of 5.920% and 4.610% over the NoEC technique, respectively. However, NoEC outperformed EC in environments Env2 (0.869) and Env3 (0.831), resulting in an average RE of 0.951. This average RE indicates a general loss of 4.890% in accuracy when using EC compared to NoEC (see Table B3).

In terms MSE for the PH trait, Table B1 shows that the use of FE achieved the best performance in most environments (15.872

[2009], 10.959 [2010], and 164.039 [2012]). However, there were exceptions in 2011 and 2013, where the best MSE values were 28.573 (EC) and 18.363 (NoEC), respectively. On the other hand, when comparing NoEC versus EC and NoEC versus FE techniques using RE, most RE values were greater than 1. On average, the use of EC and FE displayed improvements of 61.570% and 70.210%, respectively, compared to the use of NoEC. Furthermore, when comparing the performance of EC and FE techniques based on RE, FE outperformed EC in all environments, resulting in an average RE of 1.0389. This indicates that using FE surpassed EC by 3.88% (Table B1).

The EC and FE techniques were compared, using the RE metric to assess their performance. The findings indicate that the FE technique achieved its best performance in environments Env1 (1.045) and Env2 (1.048), displaying improvements of 4.480% and 4.790% over EC. However, EC exhibited a slightly better performance in environments Env3 (0.979) and Env4 (0.946), resulting in an average RE of 1.004. This average RE suggests a modest improvement of 0.430% when using FE compared to EC (see Table B3).

Predictor: E+G+GE

Table B3 reveals the performance of the FE technique in terms of MSE across different environments. The FE technique achieved its best performance in environments Env1 (2.789) and Env2 (4.636), although exceptions were found in Env3 and Env4, where the optimal MSE values were 5.833 (NoEC) and 7.792 (EC), respectively (see Table 3).

Table B3 present our comparison results between the NoEC and EC techniques, based on the RE metric. The EC technique displayed its best performance in environments Env1 (1.107) and Env4 (1.120), showing improvements of 10.72% and 12.040% over the NoEC technique. However, the NoEC technique outperformed EC in environments Env2 (0.961) and Env3 (0.925), resulting in an average RE of 1.028. This average RE indicates a general improvement of 2.840% of the EC method regarding the NoEC technique (see Table B3).

TABLE B3 The prediction performance and the relative efficiency (RE) for USP dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictors E+G and E+G+GE under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates with Feature Engineering (FE).

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G	GY	Env1	3.476	3.281	3.141	1.059	1.045	1.107
E+G	GY	Env2	4.073	4.689	4.475	0.869	1.048	0.910
E+G	GY	Env3	5.246	6.317	6.455	0.831	0.979	0.813
E+G	GY	Env4	8.174	7.814	8.262	1.046	0.946	0.989
E+G	GY	Across	–	–	–	0.951	1.004	0.955
E+G+GE	GY	Env1	3.254	2.939	2.789	1.107	1.054	1.167
E+G+GE	GY	Env2	4.708	4.898	4.636	0.961	1.057	1.016
E+G+GE	GY	Env3	5.833	6.307	6.396	0.925	0.986	0.912
E+G+GE	GY	Env4	8.730	7.792	8.206	1.120	0.950	1.064
E+G+GE	GY	Across	–	–	–	1.028	1.012	1.040

The EC and FE techniques were compared, using the RE metric to assess their performance. The findings indicate that the FE technique achieved its best performance in environments Env1 (1.054) and Env2 (1.057), displaying improvements of 5.380% and 5.650% over EC. However, using EC exhibited a better performance in environments Env3 (0.986) and Env4 (0.949), resulting in an average RE of 1.012. This average RE indicates a 1.150% improvement of the FE technique over EC (see [Table B3](#)).

Predictor: E+G+BRR

[Table B4](#) presents the results of our analysis regarding the MSE about the FE technique. The FE technique performed best in Env1 (2.859) and Env3 (4.413) environments. However, exceptions were observed in Env2 and Env4, where the optimal MSE values were 4.073 (NoEC) and 5.638 (EC), respectively. For further details, see [Table B4](#).

The results of our comparison between the NoEC and EC techniques, based on the RE metric, are presented in [Table B4](#). The EC technique exhibited its best performance in environments Env1 (1.171) and Env4 (1.450), suggesting improvements of 17.1000% and 45.000%, respectively, compared to the NoEC technique. However, the NoEC technique outperformed EC in environments Env2 (0.823) and Env3 (0.836), resulting in an average RE of 1.070. This average RE indicates a general improvement of 7.000% of the EC regarding the NoEC technique (see [Table B4](#)).

We compared the EC and FE techniques, evaluating their performance with the RE metric. The findings indicate that the FE technique achieved its best performance in environments Env1 (1.038) and Env3 (1.423), displaying respective improvements of 3.840% and 42.290% over EC. However, EC performed better in environments Env2 (0.934) and Env4 (0.990), resulting in an average RE of 1.096. This average RE indicates a 9.600% better performance of the FE technique over EC (see [Table B4](#)).

Predictor: E+G+GE+BRR

[Table B4](#) presents the performance results of the FE technique in terms of MSE. The best performance was observed in environments Env1 (2.644), Env3 (4.265), and Env4 (5.856). The only exception was Env2, where the optimal MSE value was 4.708, achieved using NoEC. For further information, see [Table B4](#).

Based on the RE metric, the results of our comparison between the NoEC and EC techniques are presented in [Table B4](#). EC performed best in environments Env1 (1.175) and Env4 (1.465), with improvements of 17.510% and 46.530%, respectively, compared to the NoEC technique. However, the NoEC technique outperformed EC in environments Env2 (0.958) and Env3 (0.915), resulting in an average RE of 1.128. This average RE indicates a general improvement of 12.830% of EC regarding NoEC. For more specific information, see [Table B4](#).

We compared the EC and FE techniques based on the RE metric. The analysis revealed that the FE technique displayed its best performance in Env1 (1.047), Env3 (1.494), and Env4 (1.017). These results indicate improvements of 4.740%, 49.430%, and 1.740%, respectively, when compared to using EC. However, EC displayed a better performance in Env2 (0.941), but in general, the FE technique outperformed EC by 12.500%, since an average RE of 1.125 was observed (see [Table B4](#)).

G2F_2016 dataset

Predictor: E+G

[Table B5](#) illustrates that FE yielded the best performance for the Grain_Moisture_BLUE trait in most environments. MSE values were 4.645 (DEH1_2016), 2.154 (GAH1_2016), 2.703 (IAH1_2016), 0.467 (IAH4_2016), 0.668 (MOH1_2016), 3.598 (NCH1_2016), 2.092 (NYH2_2016), and 1.601 (WIH2_2016). The average RE values

TABLE B4 The prediction performance and the relative efficiency (RE) for USP dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictors E+G+BRR and E+G+GE+BRR under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates with Feature Engineering (FE).

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+BRR	GY	Env1	3.476	2.968	2.859	1.171	1.038	1.216
E+G+BRR	GY	Env2	4.073	4.951	5.301	0.823	0.934	0.768
E+G+BRR	GY	Env3	5.246	6.279	4.413	0.836	1.423	1.189
E+G+BRR	GY	Env4	8.174	5.638	5.696	1.450	0.990	1.435
E+G+BRR	GY	Across	–	–	–	1.070	1.096	1.152
E+G+GE+BRR	GY	Env1	3.254	2.769	2.644	1.175	1.047	1.231
E+G+GE+BRR	GY	Env2	4.708	4.917	5.224	0.958	0.941	0.901
E+G+GE+BRR	GY	Env3	5.833	6.373	4.265	0.915	1.494	1.368
E+G+GE+BRR	GY	Env4	8.730	5.958	5.856	1.465	1.017	1.491
E+G+GE+BRR	GY	Across	–	–	–	1.128	1.125	1.248

showed that FE outperformed EC and NoEC by 87.970% and 119.370%, respectively. Additionally, EC displayed an average RE improvement of 63.960% over NoEC. For further detail, see [Table B5](#).

For the Grain_Moisture_weight trait, EC presented the best performance based on MSE values in several environments listed in [Table 5](#) (ARH1_2016 [24.235], DEH1_2016 [0.207], IAH1_2016 [2.568], ILH1_2016 [2.172], INH1_2016 [0.210], MOH1_2016 [7.450], OHH1_2016 [0.454] and WIH2_2016 [0.194]). The average RE values revealed that EC and FE outperformed the conventional NoEC technique by 1682.340% and 245.980%, respectively. Furthermore, FE displayed a 58.100% improvement over EC (See [Table B5](#)).

Regarding the Yield_Mg_ha_BLUE trait, NoEC displayed a superior performance in most environments based on MSE values listed in [Table B5](#) (GAH1_2016 [3.579], IAH4_2016 [2.576], MIH1_2016 [4.045], MNH1_2016 [1.268], NYH2_2016 [16.252], OHH1_2016 [1.830] and WIH1_2016 [3.665]). The average RE values indicated that FE resulted in general improvements of 21.030% and 1.400% over EC and NoEC, respectively. However, a comparison between NoEC and EC showed a slight decrease of 0.190% in average RE for EC (see [Table B5](#)).

For the Yield_Mg_ha_weight trait, NoEC showed the best performance based on MSE values in most environments (DEH1_2016 [0.078], IAH4_2016 [0.091], ILH1_2016 [0.351], MIH1_2016 [0.1156], MNH1_2016 [0.391], NYH2_2016 [0.087], WIH1_2016 [0.063] and WIH2_2016 [0.019]). The average RE values indicated general improvements of 52.860% and 22.630% for EC and FE, respectively, compared to NoEC. Moreover, on average, FE outperformed EC by 89.600% (see [Table B5](#)).

Predictor: E+G+GE

[Table B5](#) shows that FE yielded the best performance for the Grain_Moisture_BLUE trait in the majority of environments, with MSE values ranging from 0.519 to 5.813 (IAH4_2016, ILH1_2016, MNH1_2016, NEH1_2016, NYH2_2016, OHH1_2016 and WIH1_2016). Comparing RE values, using FE outperformed EC and NoEC techniques by 42.480% and 114.740%, respectively. Additionally, EC outperformed NoEC with an average RE of 1.552, indicating a superiority of 55.210% for EC. For further details, see [Table B5](#).

For the Grain_Moisture_weight trait, [Table B5](#) reveals that FE displayed a better performance in most environments, as indicated by the MSE values (DEH1_2016 [0.132], IAH3_2016 [0.418], IAH4_2016 [139.446], MIH1_2016 [1.668], MNH1_2016 [1.316], NCH1_2016 [6.953], NYH2_2016 [5.565], OHH1_2016 [0.195] and WIH1_2016 [1.508]). Moreover, the average RE values showed that FE outperformed EC and NoEC by 831.910% and 825.260%, respectively. Comparing NoEC and EC techniques, there was a general improvement of 357.000% for EC over NoEC, with an average RE of 4.570 (see [Table B5](#)).

Regarding the Yield_Mg_ha_BLUE trait, [Table B5](#) shows that the use of NoEC achieved the best performance in most environments, as indicated by the MSE values (GAH1_2016 [3.379], IAH1_2016 [2.287], IAH2_2016 [7.505], IAH4_2016 [3.565], MIH1_2016 [4.748], NYH2_2016 [17.271], WIH1_2016 [2.210] and WIH2_2016 [4.667]). However, most RE values comparing NoEC_vs_EC and NoEC_vs_FENoEC_vs_FE techniques were greater than 1. On average, EC displayed a 7.450% improvement and FE showed an 11.690% improvement compared to the

TABLE B5 The prediction performance and the relative efficiency (RE) for G2F_2016 dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictors E+G and E+G+GE under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates with Feature Engineering (FE).

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G	Grain_Moisture_BLUE	ARH1_2016	1.733	6.108	1.886	0.284	3.238	0.919
E+G	Grain_Moisture_BLUE	DEH1_2016	7.863	5.829	4.645	1.349	1.255	1.693
E+G	Grain_Moisture_BLUE	GAH1_2016	6.686	5.107	2.154	1.309	2.371	3.105
E+G	Grain_Moisture_BLUE	IAH1_2016	9.814	7.419	2.703	1.323	2.745	3.632
E+G	Grain_Moisture_BLUE	IAH2_2016	3.124	0.866	1.694	3.608	0.511	1.844
E+G	Grain_Moisture_BLUE	IAH3_2016	1.456	2.981	1.486	0.489	2.006	0.980
E+G	Grain_Moisture_BLUE	IAH4_2016	2.495	0.506	0.467	4.932	1.084	5.344
E+G	Grain_Moisture_BLUE	ILH1_2016	4.556	3.436	9.783	1.326	0.351	0.466
E+G	Grain_Moisture_BLUE	INH1_2016	1.934	9.982	2.887	0.194	3.457	0.670
E+G	Grain_Moisture_BLUE	MIH1_2016	2.988	3.101	3.366	0.963	0.922	0.888
E+G	Grain_Moisture_BLUE	MNH1_2016	17.117	4.471	4.483	3.829	0.997	3.818
E+G	Grain_Moisture_BLUE	MOH1_2016	0.809	3.068	0.668	0.264	4.593	1.211
E+G	Grain_Moisture_BLUE	NCH1_2016	21.208	10.860	3.598	1.953	3.018	5.895
E+G	Grain_Moisture_BLUE	NEH1_2016	6.193	4.897	10.060	1.265	0.487	0.616

(Continued)

TABLE B5 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G	Grain_Moisture_BLUE	NYH2_2016	7.475	3.625	2.092	2.062	1.732	3.573
E+G	Grain_Moisture_BLUE	OHH1_2016	4.840	2.834	5.728	1.708	0.495	0.845
E+G	Grain_Moisture_BLUE	WIH1_2016	5.143	2.599	3.788	1.979	0.686	1.358
E+G	Grain_Moisture_BLUE	WIH2_2016	4.219	6.224	1.601	0.678	3.887	2.634
E+G	Grain_Moisture_BLUE	Across	–	–	–	1.640	1.880	2.194
E+G	Grain_Moisture_weight	ARH1_2016	30.391	24.235	30.934	1.254	0.783	0.982
E+G	Grain_Moisture_weight	DEH1_2016	14.987	0.207	2.910	72.261	0.071	5.150
E+G	Grain_Moisture_weight	GAH1_2016	1.272	2.339	7.133	0.544	0.328	0.178
E+G	Grain_Moisture_weight	IAH1_2016	401.574	481.573	510.263	0.834	0.944	0.787
E+G	Grain_Moisture_weight	IAH2_2016	6.212	2.568	25.510	2.419	0.101	0.244
E+G	Grain_Moisture_weight	IAH3_2016	0.199	10.913	31.831	0.018	0.343	0.006
E+G	Grain_Moisture_weight	IAH4_2016	311.023	244.775	180.018	1.271	1.360	1.728
E+G	Grain_Moisture_weight	ILH1_2016	5.447	2.172	25.900	2.507	0.084	0.210
E+G	Grain_Moisture_weight	INH1_2016	1.274	0.210	0.325	6.058	0.647	3.916
E+G	Grain_Moisture_weight	MIH1_2016	0.715	8.311	0.872	0.086	9.531	0.820
E+G	Grain_Moisture_weight	MNH1_2016	7.866	43.427	6.379	0.181	6.808	1.233
E+G	Grain_Moisture_weight	MOH1_2016	27.122	7.450	44.113	3.640	0.169	0.615
E+G	Grain_Moisture_weight	NCH1_2016	1.174	4.278	10.042	0.274	0.426	0.117
E+G	Grain_Moisture_weight	NEH1_2016	42.758	63.944	64.869	0.669	0.986	0.659
E+G	Grain_Moisture_weight	NYH2_2016	1.893	2.551	11.545	0.742	0.221	0.164
E+G	Grain_Moisture_weight	OHH1_2016	63.776	0.454	27.250	140.383	0.017	2.340
E+G	Grain_Moisture_weight	WIH1_2016	1.373	7.073	1.371	0.194	5.160	1.001
E+G	Grain_Moisture_weight	WIH2_2016	16.972	0.194	0.403	87.485	0.482	42.125
E+G	Grain_Moisture_weight	Across	–	–	–	17.823	1.581	3.460
E+G	Yield_Mg_ha_BLUE	ARH1_2016	3.713	3.199	14.552	1.161	0.220	0.255
E+G	Yield_Mg_ha_BLUE	DEH1_2016	5.330	3.354	4.354	1.589	0.770	1.224
E+G	Yield_Mg_ha_BLUE	GAH1_2016	3.580	10.264	4.606	0.349	2.229	0.777
E+G	Yield_Mg_ha_BLUE	IAH1_2016	3.187	2.897	1.395	1.100	2.077	2.286
E+G	Yield_Mg_ha_BLUE	IAH2_2016	7.921	7.684	8.073	1.031	0.952	0.981
E+G	Yield_Mg_ha_BLUE	IAH3_2016	5.918	4.741	3.772	1.248	1.257	1.569
E+G	Yield_Mg_ha_BLUE	IAH4_2016	2.576	2.718	3.708	0.948	0.733	0.695
E+G	Yield_Mg_ha_BLUE	ILH1_2016	8.719	4.698	6.260	1.856	0.750	1.393
E+G	Yield_Mg_ha_BLUE	INH1_2016	2.415	3.018	2.406	0.800	1.254	1.004
E+G	Yield_Mg_ha_BLUE	MIH1_2016	4.045	5.627	16.686	0.719	0.337	0.242
E+G	Yield_Mg_ha_BLUE	MNH1_2016	1.268	1.301	1.270	0.975	1.025	0.999
E+G	Yield_Mg_ha_BLUE	MOH1_2016	7.968	4.191	10.428	1.901	0.402	0.764
E+G	Yield_Mg_ha_BLUE	NCH1_2016	4.467	10.571	3.293	0.423	3.211	1.357
E+G	Yield_Mg_ha_BLUE	NEH1_2016	4.993	4.832	4.188	1.033	1.154	1.192
E+G	Yield_Mg_ha_BLUE	NYH2_2016	16.252	22.790	16.626	0.713	1.371	0.978

(Continued)

TABLE B5 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G	Yield_Mg_ha_BLUE	OHH1_2016	1.830	4.790	2.558	0.382	1.872	0.715
E+G	Yield_Mg_ha_BLUE	WIH1_2016	3.665	5.021	3.785	0.730	1.326	0.968
E+G	Yield_Mg_ha_BLUE	WIH2_2016	4.630	4.588	5.420	1.009	0.846	0.854
E+G	Yield_Mg_ha_BLUE	Across	–	–	–	0.998	1.210	1.014
E+G	Yield_Mg_ha_weight	ARH1_2016	0.989	1.000	1.542	0.989	0.649	0.641
E+G	Yield_Mg_ha_weight	DEH1_2016	0.163	0.078	0.230	2.088	0.340	0.710
E+G	Yield_Mg_ha_weight	GAH1_2016	0.035	0.439	0.288	0.079	1.522	0.120
E+G	Yield_Mg_ha_weight	IAH1_2016	3.743	3.345	3.151	1.119	1.061	1.188
E+G	Yield_Mg_ha_weight	IAH2_2016	0.175	0.668	0.077	0.262	8.629	2.261
E+G	Yield_Mg_ha_weight	IAH3_2016	0.788	1.704	1.583	0.462	1.076	0.498
E+G	Yield_Mg_ha_weight	IAH4_2016	0.498	0.091	0.105	5.491	0.861	4.729
E+G	Yield_Mg_ha_weight	ILH1_2016	1.113	0.351	0.754	3.172	0.465	1.476
E+G	Yield_Mg_ha_weight	INH1_2016	0.055	0.077	0.052	0.709	1.486	1.054
E+G	Yield_Mg_ha_weight	MIH1_2016	0.121	0.116	0.132	1.042	0.875	0.912
E+G	Yield_Mg_ha_weight	MNH1_2016	0.393	0.391	0.711	1.005	0.551	0.553
E+G	Yield_Mg_ha_weight	MOH1_2016	0.232	1.501	0.172	0.155	8.721	1.348
E+G	Yield_Mg_ha_weight	NCH1_2016	0.083	0.343	0.085	0.241	4.062	0.978
E+G	Yield_Mg_ha_weight	NEH1_2016	0.036	0.038	0.029	0.963	1.279	1.231
E+G	Yield_Mg_ha_weight	NYH2_2016	0.402	0.087	0.139	4.601	0.630	2.899
E+G	Yield_Mg_ha_weight	OHH1_2016	0.533	1.326	0.876	0.402	1.514	0.608
E+G	Yield_Mg_ha_weight	WIH1_2016	0.117	0.063	0.209	1.877	0.299	0.561
E+G	Yield_Mg_ha_weight	WIH2_2016	0.055	0.019	0.180	2.860	0.107	0.306
E+G	Yield_Mg_ha_weight	Across	–	–	–	1.529	1.896	1.226
E+G+GE	Grain_Moisture_BLUE	ARH1_2016	2.003	6.545	4.641	0.306	1.410	0.432
E+G+GE	Grain_Moisture_BLUE	DEH1_2016	5.256	5.689	10.400	0.924	0.547	0.505
E+G+GE	Grain_Moisture_BLUE	GAH1_2016	5.841	3.993	2.715	1.463	1.471	2.152
E+G+GE	Grain_Moisture_BLUE	IAH1_2016	2.857	5.585	3.541	0.512	1.577	0.807
E+G+GE	Grain_Moisture_BLUE	IAH2_2016	0.713	1.504	1.785	0.475	0.843	0.400
E+G+GE	Grain_Moisture_BLUE	IAH3_2016	2.933	4.648	2.860	0.631	1.625	1.025
E+G+GE	Grain_Moisture_BLUE	IAH4_2016	1.622	0.519	0.695	3.123	0.747	2.333
E+G+GE	Grain_Moisture_BLUE	ILH1_2016	8.071	4.093	9.622	1.972	0.425	0.839
E+G+GE	Grain_Moisture_BLUE	INH1_2016	5.315	10.531	4.891	0.505	2.153	1.087
E+G+GE	Grain_Moisture_BLUE	MIH1_2016	2.448	3.313	5.501	0.739	0.602	0.445
E+G+GE	Grain_Moisture_BLUE	MNH1_2016	13.571	5.813	6.414	2.335	0.906	2.116
E+G+GE	Grain_Moisture_BLUE	MOH1_2016	3.450	5.296	1.357	0.651	3.904	2.543
E+G+GE	Grain_Moisture_BLUE	NCH1_2016	14.869	8.231	2.333	1.806	3.528	6.374
E+G+GE	Grain_Moisture_BLUE	NEH1_2016	12.527	5.166	10.466	2.425	0.494	1.197
E+G+GE	Grain_Moisture_BLUE	NYH2_2016	9.727	4.423	5.172	2.199	0.855	1.881
E+G+GE	Grain_Moisture_BLUE	OHH1_2016	6.975	2.849	6.176	2.448	0.461	1.129

(Continued)

TABLE B5 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE	Grain_Moisture_BLUE	WIH1_2016	6.024	3.056	5.975	1.971	0.512	1.008
E+G+GE	Grain_Moisture_BLUE	WIH2_2016	21.532	6.235	1.739	3.454	3.585	12.382
E+G+GE	Grain_Moisture_BLUE	Across	–	–	–	1.552	1.425	2.147
E+G+GE	Grain_Moisture_weight	ARH1_2016	14.116	3.206	9.123	4.403	0.351	1.547
E+G+GE	Grain_Moisture_weight	DEH1_2016	1.608	10.772	0.132	0.149	81.919	12.231
E+G+GE	Grain_Moisture_weight	GAH1_2016	0.862	0.883	4.521	0.976	0.195	0.191
E+G+GE	Grain_Moisture_weight	IAH1_2016	501.269	514.108	546.300	0.975	0.941	0.918
E+G+GE	Grain_Moisture_weight	IAH2_2016	43.354	23.631	36.310	1.835	0.651	1.194
E+G+GE	Grain_Moisture_weight	IAH3_2016	11.456	7.015	0.418	1.633	16.769	27.387
E+G+GE	Grain_Moisture_weight	IAH4_2016	265.697	167.322	139.446	1.588	1.200	1.905
E+G+GE	Grain_Moisture_weight	ILH1_2016	35.818	2.973	32.902	12.047	0.090	1.089
E+G+GE	Grain_Moisture_weight	INH1_2016	51.327	1.919	3.812	26.741	0.504	13.465
E+G+GE	Grain_Moisture_weight	MIH1_2016	18.430	38.977	1.668	0.473	23.368	11.049
E+G+GE	Grain_Moisture_weight	MNH1_2016	11.304	39.937	1.316	0.283	30.345	8.589
E+G+GE	Grain_Moisture_weight	MOH1_2016	3.665	14.395	291.204	0.255	0.049	0.013
E+G+GE	Grain_Moisture_weight	NCH1_2016	7.758	7.873	6.953	0.985	1.132	1.116
E+G+GE	Grain_Moisture_weight	NEH1_2016	113.669	88.519	99.451	1.284	0.890	1.143
E+G+GE	Grain_Moisture_weight	NYH2_2016	80.595	16.174	5.565	4.983	2.906	14.482
E+G+GE	Grain_Moisture_weight	OHH1_2016	12.108	0.596	0.195	20.319	3.054	62.060
E+G+GE	Grain_Moisture_weight	WIH1_2016	11.902	4.475	1.508	2.660	2.967	7.892
E+G+GE	Grain_Moisture_weight	WIH2_2016	0.917	1.365	3.320	0.672	0.411	0.276
E+G+GE	Grain_Moisture_weight	Across	–	–	–	4.570	9.319	9.253
E+G+GE	Yield_Mg_ha_BLUE	ARH1_2016	3.928	2.896	14.301	1.357	0.203	0.275
E+G+GE	Yield_Mg_ha_BLUE	DEH1_2016	5.964	3.522	3.831	1.694	0.919	1.557
E+G+GE	Yield_Mg_ha_BLUE	GAH1_2016	3.379	10.667	4.157	0.317	2.566	0.813
E+G+GE	Yield_Mg_ha_BLUE	IAH1_2016	2.287	2.778	2.820	0.823	0.985	0.811
E+G+GE	Yield_Mg_ha_BLUE	IAH2_2016	7.505	8.311	7.733	0.903	1.075	0.971
E+G+GE	Yield_Mg_ha_BLUE	IAH3_2016	7.908	6.619	5.280	1.195	1.254	1.498
E+G+GE	Yield_Mg_ha_BLUE	IAH4_2016	2.565	2.895	3.811	0.886	0.760	0.673
E+G+GE	Yield_Mg_ha_BLUE	ILH1_2016	8.036	4.761	5.919	1.688	0.804	1.358
E+G+GE	Yield_Mg_ha_BLUE	INH1_2016	6.533	2.424	1.994	2.696	1.216	3.277
E+G+GE	Yield_Mg_ha_BLUE	MIH1_2016	4.748	7.252	19.667	0.655	0.369	0.241
E+G+GE	Yield_Mg_ha_BLUE	MNH1_2016	1.422	1.479	1.265	0.961	1.169	1.124
E+G+GE	Yield_Mg_ha_BLUE	MOH1_2016	12.381	5.928	9.392	2.089	0.631	1.318
E+G+GE	Yield_Mg_ha_BLUE	NCH1_2016	5.713	11.008	3.515	0.519	3.132	1.626
E+G+GE	Yield_Mg_ha_BLUE	NEH1_2016	5.446	5.707	5.214	0.954	1.095	1.045
E+G+GE	Yield_Mg_ha_BLUE	NYH2_2016	17.271	24.594	19.504	0.702	1.261	0.886
E+G+GE	Yield_Mg_ha_BLUE	OHH1_2016	2.503	4.763	2.138	0.526	2.227	1.171
E+G+GE	Yield_Mg_ha_BLUE	WIH1_2016	2.210	5.855	3.805	0.378	1.539	0.581

(Continued)

TABLE B5 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE	Yield_Mg_ha_BLUE	WIH2_2016	4.667	4.667	5.288	1.000	0.883	0.883
E+G+GE	Yield_Mg_ha_BLUE	Across	–	–	–	1.075	1.227	1.117
E+G+GE	Yield_Mg_ha_weight	ARH1_2016	2.359	1.339	1.540	1.762	0.869	1.532
E+G+GE	Yield_Mg_ha_weight	DEH1_2016	0.051	0.124	0.284	0.410	0.437	0.179
E+G+GE	Yield_Mg_ha_weight	GAH1_2016	0.026	0.357	0.258	0.074	1.385	0.102
E+G+GE	Yield_Mg_ha_weight	IAH1_2016	2.914	3.540	3.508	0.823	1.009	0.831
E+G+GE	Yield_Mg_ha_weight	IAH2_2016	0.069	0.410	0.076	0.168	5.378	0.903
E+G+GE	Yield_Mg_ha_weight	IAH3_2016	0.670	0.608	1.186	1.102	0.513	0.565
E+G+GE	Yield_Mg_ha_weight	IAH4_2016	0.199	0.110	0.082	1.807	1.343	2.426
E+G+GE	Yield_Mg_ha_weight	ILH1_2016	0.751	0.468	0.539	1.605	0.868	1.394
E+G+GE	Yield_Mg_ha_weight	INH1_2016	0.112	0.056	0.046	1.981	1.227	2.429
E+G+GE	Yield_Mg_ha_weight	MIH1_2016	0.055	0.189	0.172	0.291	1.098	0.320
E+G+GE	Yield_Mg_ha_weight	MNH1_2016	0.146	0.352	0.502	0.415	0.701	0.291
E+G+GE	Yield_Mg_ha_weight	MOH1_2016	0.283	0.295	0.263	0.959	1.122	1.076
E+G+GE	Yield_Mg_ha_weight	NCH1_2016	0.113	0.388	0.104	0.292	3.730	1.090
E+G+GE	Yield_Mg_ha_weight	NEH1_2016	0.033	0.073	0.081	0.458	0.900	0.412
E+G+GE	Yield_Mg_ha_weight	NYH2_2016	0.449	0.781	0.709	0.575	1.102	0.633
E+G+GE	Yield_Mg_ha_weight	OHH1_2016	1.202	1.667	1.328	0.721	1.255	0.905
E+G+GE	Yield_Mg_ha_weight	WIH1_2016	0.204	0.096	0.132	2.125	0.729	1.550
E+G+GE	Yield_Mg_ha_weight	WIH2_2016	0.185	0.091	0.204	2.036	0.443	0.903
E+G+GE	Yield_Mg_ha_weight	Across	–	–	–	0.978	1.339	0.974

conventional NoEC technique. Furthermore, comparing EC and FE techniques, an average RE of 1.227 was observed, indicating that FE outperformed NoEC by 22.700% (see [Table B5](#)).

In terms of the Yield_Mg_ha_weight trait, [Table B5](#) shows that the use of NoEC achieved the best performance in most environments, as evident from the MSE values (DEH1_2016 [0.051], GAH1_2016 [0.026], IAH1_2016 [2.914], IAH2_2016 [0.0689], MIH1_2016 [0.055], MNH1_2016 [0.146], NEH1_2016 [0.033], NYH2_2016 [0.449] and OHH1_2016 [1.202]). The average RE values indicated slight losses of 2.210% and 2.570% when comparing EC versus NoEC and FE versus NoEC, respectively. This implies that EC and FE techniques did not perform as adequately as the conventional NoEC technique. However, comparing EC and FE techniques based on RE showed that FE outperformed EC in most environments, resulting in an average RE of 1.339, indicating a 33.930% superiority of FE over EC. For more detailed information, see [Table B5](#).

Predictor: E+G+BRR

In [Table B6](#), it is evident that for the Grain_Moisture_BLUE trait, the use of FE provided the best performance in most environments, as indicated by the MSE values (DEH1_2016

[4.376], GAH1_2016 [2.002], IAH1_2016 [2.036], IAH3_2016 [1.237], IAH4_2016 [0.496], MNH1_2016 [3.685], MOH1_2016 [0.678], NCH1_2016 [3.499], NYH2_2016 [2.213] and WIH2_2016 [1.648]). On average, the RE values indicate that FE outperformed EC and NoEC by 67.090% and 129.850%, respectively. Additionally, comparing NoEC and EC techniques showed that EC outperformed NoEC by an average of 84.880%. For further information, see [Table B6](#).

For the Grain_Moisture_weight trait, [Table B6](#) shows that the use of NoEC provided the best performance in most environments, as indicated by the MSE values (GAH1_2016 [1.272], IAH1_2016 [401.574], IAH3_2016 [0.199], ILH1_2016 [5.447], MIH1_2016 [0.715], NCH1_2016 [1.174] and NEH1_2016 [42.758]). On average, the RE values indicate that FE outperformed EC and NoEC by 167.270% and 25.410%, respectively. Furthermore, comparing NoEC and EC shows that EC outperformed NoEC with an average RE of 3.495, representing a general improvement of 149.510%. For more detailed information, see [Table B6](#).

[Table B6](#), for the Yield_Mg_ha_BLUE trait, shows that the use of NoEC led to the best performance in most environments, as indicated by the MSE values (ARH1_2016 [3.713], GAH1_2016 [3.579], IAH4_2016 [2.576], INH1_2016 [2016], MIH1_2016

TABLE B6 The prediction performance and the relative efficiency (RE) for G2F_2016 dataset in terms of mean squared error (MSE) for each Environment and for each trait, for the predictor E+G+BRR and E+G+GE+BRR under three different techniques to compute the Kernel for the effect of the Environment: without Environmental Covariates (NoEC), using Environmental covariates (EC) and using Environmental Covariates with Feature Engineering (FE).

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+BRR	Grain_Moisture_BLUE	ARH1_2016	1.733	8.086	2.856	0.214	2.832	0.607
E+G+BRR	Grain_Moisture_BLUE	DEH1_2016	7.863	5.151	4.376	1.526	1.177	1.797
E+G+BRR	Grain_Moisture_BLUE	GAH1_2016	6.686	5.025	2.002	1.331	2.511	3.341
E+G+BRR	Grain_Moisture_BLUE	IAH1_2016	9.814	3.372	2.036	2.911	1.656	4.821
E+G+BRR	Grain_Moisture_BLUE	IAH2_2016	3.124	1.172	1.650	2.665	0.711	1.894
E+G+BRR	Grain_Moisture_BLUE	IAH3_2016	1.456	2.060	1.237	0.707	1.665	1.178
E+G+BRR	Grain_Moisture_BLUE	IAH4_2016	2.495	0.515	0.496	4.845	1.039	5.031
E+G+BRR	Grain_Moisture_BLUE	ILH1_2016	4.556	2.970	9.745	1.534	0.305	0.468
E+G+BRR	Grain_Moisture_BLUE	INH1_2016	1.934	12.122	2.526	0.160	4.798	0.766
E+G+BRR	Grain_Moisture_BLUE	MIH1_2016	2.988	3.562	3.335	0.839	1.068	0.896
E+G+BRR	Grain_Moisture_BLUE	MNH1_2016	17.117	3.852	3.685	4.444	1.045	4.645
E+G+BRR	Grain_Moisture_BLUE	MOH1_2016	0.809	1.362	0.678	0.594	2.009	1.193
E+G+BRR	Grain_Moisture_BLUE	NCH1_2016	21.208	10.060	3.499	2.108	2.875	6.061
E+G+BRR	Grain_Moisture_BLUE	NEH1_2016	6.193	2.846	9.795	2.176	0.291	0.632
E+G+BRR	Grain_Moisture_BLUE	NYH2_2016	7.475	2.344	2.213	3.189	1.059	3.378
E+G+BRR	Grain_Moisture_BLUE	OHH1_2016	4.840	2.898	5.870	1.670	0.494	0.825
E+G+BRR	Grain_Moisture_BLUE	WIH1_2016	5.143	3.045	4.014	1.689	0.759	1.281
E+G+BRR	Grain_Moisture_BLUE	WIH2_2016	4.219	6.235	1.648	0.677	3.785	2.560
E+G+BRR	Grain_Moisture_BLUE	Across	–	–	–	1.849	1.671	2.299
E+G+BRR	Grain_Moisture_weight	ARH1_2016	30.391	7.962	7.088	3.817	1.123	4.288
E+G+BRR	Grain_Moisture_weight	DEH1_2016	14.987	0.443	5.442	33.869	0.081	2.754
E+G+BRR	Grain_Moisture_weight	GAH1_2016	1.272	5.393	2.233	0.236	2.415	0.569
E+G+BRR	Grain_Moisture_weight	IAH1_2016	401.574	459.125	508.319	0.875	0.903	0.790
E+G+BRR	Grain_Moisture_weight	IAH2_2016	6.212	1.611	176.584	3.855	0.009	0.035
E+G+BRR	Grain_Moisture_weight	IAH3_2016	0.199	51.438	110.303	0.004	0.466	0.002
E+G+BRR	Grain_Moisture_weight	IAH4_2016	311.023	188.044	160.261	1.654	1.173	1.941
E+G+BRR	Grain_Moisture_weight	ILH1_2016	5.447	22.946	64.425	0.237	0.356	0.085
E+G+BRR	Grain_Moisture_weight	INH1_2016	1.274	0.691	0.685	1.843	1.009	1.860
E+G+BRR	Grain_Moisture_weight	MIH1_2016	0.715	31.083	1.554	0.023	20.002	0.460
E+G+BRR	Grain_Moisture_weight	MNH1_2016	7.866	43.882	6.124	0.179	7.165	1.284
E+G+BRR	Grain_Moisture_weight	MOH1_2016	27.122	21.394	393.212	1.268	0.054	0.069
E+G+BRR	Grain_Moisture_weight	NCH1_2016	1.174	5.985	24.041	0.196	0.249	0.049
E+G+BRR	Grain_Moisture_weight	NEH1_2016	42.758	57.295	90.340	0.746	0.634	0.473
E+G+BRR	Grain_Moisture_weight	NYH2_2016	1.893	0.666	46.015	2.842	0.015	0.041
E+G+BRR	Grain_Moisture_weight	OHH1_2016	63.776	7.228	19.206	8.823	0.376	3.321
E+G+BRR	Grain_Moisture_weight	WIH1_2016	1.373	13.412	1.266	0.102	10.595	1.084
E+G+BRR	Grain_Moisture_weight	WIH2_2016	16.972	7.246	4.891	2.342	1.482	3.470

(Continued)

TABLE B6 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+BRR	Grain_Moisture_weight	Across	–	–	–	3.495	2.673	1.254
E+G+BRR	Yield_Mg_ha_BLUE	ARH1_2016	3.713	3.799	14.569	0.977	0.261	0.255
E+G+BRR	Yield_Mg_ha_BLUE	DEH1_2016	5.330	2.922	3.946	1.824	0.740	1.351
E+G+BRR	Yield_Mg_ha_BLUE	GAH1_2016	3.580	11.055	5.613	0.324	1.970	0.638
E+G+BRR	Yield_Mg_ha_BLUE	IAH1_2016	3.187	1.743	1.393	1.829	1.252	2.289
E+G+BRR	Yield_Mg_ha_BLUE	IAH2_2016	7.921	7.568	8.528	1.047	0.888	0.929
E+G+BRR	Yield_Mg_ha_BLUE	IAH3_2016	5.918	5.873	6.247	1.008	0.940	0.947
E+G+BRR	Yield_Mg_ha_BLUE	IAH4_2016	2.576	2.618	3.773	0.984	0.694	0.683
E+G+BRR	Yield_Mg_ha_BLUE	ILH1_2016	8.719	4.687	7.329	1.860	0.640	1.190
E+G+BRR	Yield_Mg_ha_BLUE	INH1_2016	2.415	2.675	2.435	0.903	1.098	0.992
E+G+BRR	Yield_Mg_ha_BLUE	MIH1_2016	4.045	6.342	17.412	0.638	0.364	0.232
E+G+BRR	Yield_Mg_ha_BLUE	MNH1_2016	1.268	1.350	1.270	0.939	1.063	0.999
E+G+BRR	Yield_Mg_ha_BLUE	MOH1_2016	7.968	4.093	10.724	1.947	0.382	0.743
E+G+BRR	Yield_Mg_ha_BLUE	NCH1_2016	4.467	9.870	3.889	0.453	2.538	1.149
E+G+BRR	Yield_Mg_ha_BLUE	NEH1_2016	4.993	4.703	3.515	1.062	1.338	1.421
E+G+BRR	Yield_Mg_ha_BLUE	NYH2_2016	16.252	22.892	17.091	0.710	1.339	0.951
E+G+BRR	Yield_Mg_ha_BLUE	OHH1_2016	1.830	4.374	2.456	0.418	1.781	0.745
E+G+BRR	Yield_Mg_ha_BLUE	WIH1_2016	3.665	4.548	2.558	0.806	1.778	1.433
E+G+BRR	Yield_Mg_ha_BLUE	WIH2_2016	4.630	4.859	5.700	0.953	0.853	0.812
E+G+BRR	Yield_Mg_ha_BLUE	Across	–	–	–	1.038	1.107	0.986
E+G+BRR	Yield_Mg_ha_weight	ARH1_2016	0.989	1.219	1.311	0.811	0.930	0.755
E+G+BRR	Yield_Mg_ha_weight	DEH1_2016	0.163	0.029	0.076	5.723	0.375	2.143
E+G+BRR	Yield_Mg_ha_weight	GAH1_2016	0.035	0.251	0.134	0.138	1.870	0.259
E+G+BRR	Yield_Mg_ha_weight	IAH1_2016	3.743	3.506	3.050	1.068	1.150	1.227
E+G+BRR	Yield_Mg_ha_weight	IAH2_2016	0.175	0.372	3.081	0.471	0.121	0.057
E+G+BRR	Yield_Mg_ha_weight	IAH3_2016	0.788	0.976	2.401	0.808	0.407	0.328
E+G+BRR	Yield_Mg_ha_weight	IAH4_2016	0.498	0.065	0.179	7.678	0.362	2.782
E+G+BRR	Yield_Mg_ha_weight	ILH1_2016	1.113	0.336	0.581	3.316	0.578	1.916
E+G+BRR	Yield_Mg_ha_weight	INH1_2016	0.055	0.044	0.058	1.239	0.761	0.943
E+G+BRR	Yield_Mg_ha_weight	MIH1_2016	0.121	0.297	0.300	0.406	0.992	0.402
E+G+BRR	Yield_Mg_ha_weight	MNH1_2016	0.393	0.721	0.682	0.546	1.057	0.577
E+G+BRR	Yield_Mg_ha_weight	MOH1_2016	0.232	0.521	0.252	0.445	2.066	0.920
E+G+BRR	Yield_Mg_ha_weight	NCH1_2016	0.083	0.311	0.078	0.266	4.012	1.066
E+G+BRR	Yield_Mg_ha_weight	NEH1_2016	0.036	0.030	0.031	1.203	0.984	1.183
E+G+BRR	Yield_Mg_ha_weight	NYH2_2016	0.402	0.419	0.700	0.960	0.598	0.574
E+G+BRR	Yield_Mg_ha_weight	OHH1_2016	0.533	1.561	1.276	0.341	1.224	0.418
E+G+BRR	Yield_Mg_ha_weight	WIH1_2016	0.117	0.067	0.207	1.746	0.324	0.566
E+G+BRR	Yield_Mg_ha_weight	WIH2_2016	0.055	0.424	0.469	0.130	0.904	0.118
E+G+BRR	Yield_Mg_ha_weight	Across	–	–	–	1.516	1.040	0.902

(Continued)

TABLE B6 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE+BRR	Grain_Moisture_BLUE	ARH1_2016	2.003	8.861	8.335	0.226	1.063	0.240
E+G+GE+BRR	Grain_Moisture_BLUE	DEH1_2016	5.256	4.281	9.799	1.228	0.437	0.536
E+G+GE+BRR	Grain_Moisture_BLUE	GAH1_2016	5.841	4.396	2.596	1.329	1.693	2.250
E+G+GE+BRR	Grain_Moisture_BLUE	IAH1_2016	2.857	3.613	2.833	0.791	1.275	1.008
E+G+GE+BRR	Grain_Moisture_BLUE	IAH2_2016	0.713	1.881	1.708	0.379	1.101	0.418
E+G+GE+BRR	Grain_Moisture_BLUE	IAH3_2016	2.933	3.283	2.610	0.893	1.258	1.124
E+G+GE+BRR	Grain_Moisture_BLUE	IAH4_2016	1.622	0.519	0.724	3.127	0.717	2.241
E+G+GE+BRR	Grain_Moisture_BLUE	ILH1_2016	8.071	4.964	9.657	1.626	0.514	0.836
E+G+GE+BRR	Grain_Moisture_BLUE	INH1_2016	5.315	11.384	4.258	0.467	2.674	1.248
E+G+GE+BRR	Grain_Moisture_BLUE	MIH1_2016	2.448	3.737	3.645	0.655	1.025	0.672
E+G+GE+BRR	Grain_Moisture_BLUE	MNH1_2016	13.571	4.762	6.621	2.850	0.719	2.050
E+G+GE+BRR	Grain_Moisture_BLUE	MOH1_2016	3.450	2.107	1.221	1.637	1.725	2.825
E+G+GE+BRR	Grain_Moisture_BLUE	NCH1_2016	14.869	7.756	2.226	1.917	3.485	6.681
E+G+GE+BRR	Grain_Moisture_BLUE	NEH1_2016	12.527	6.047	10.506	2.072	0.576	1.192
E+G+GE+BRR	Grain_Moisture_BLUE	NYH2_2016	9.727	4.030	5.378	2.414	0.749	1.809
E+G+GE+BRR	Grain_Moisture_BLUE	OHH1_2016	6.975	3.072	8.466	2.270	0.363	0.824
E+G+GE+BRR	Grain_Moisture_BLUE	WIH1_2016	6.024	3.495	6.151	1.723	0.568	0.979
E+G+GE+BRR	Grain_Moisture_BLUE	WIH2_2016	21.532	5.216	1.584	4.128	3.293	13.594
E+G+GE+BRR	Grain_Moisture_BLUE	Across	–	–	–	1.652	1.291	2.252
E+G+GE+BRR	Grain_Moisture_weight	ARH1_2016	14.116	33.005	48.706	0.428	0.678	0.290
E+G+GE+BRR	Grain_Moisture_weight	DEH1_2016	1.608	0.595	0.683	2.701	0.872	2.355
E+G+GE+BRR	Grain_Moisture_weight	GAH1_2016	0.862	3.261	1.258	0.264	2.593	0.685
E+G+GE+BRR	Grain_Moisture_weight	IAH1_2016	501.269	360.363	452.522	1.391	0.796	1.108
E+G+GE+BRR	Grain_Moisture_weight	IAH2_2016	43.354	1.219	28.797	35.562	0.042	1.506
E+G+GE+BRR	Grain_Moisture_weight	IAH3_2016	11.456	92.472	220.035	0.124	0.420	0.052
E+G+GE+BRR	Grain_Moisture_weight	IAH4_2016	265.697	120.354	139.962	2.208	0.860	1.898
E+G+GE+BRR	Grain_Moisture_weight	ILH1_2016	35.818	10.357	65.451	3.459	0.158	0.547
E+G+GE+BRR	Grain_Moisture_weight	INH1_2016	51.327	29.589	16.709	1.735	1.771	3.072
E+G+GE+BRR	Grain_Moisture_weight	MIH1_2016	18.430	47.158	9.360	0.391	5.039	1.969
E+G+GE+BRR	Grain_Moisture_weight	MNH1_2016	11.304	52.703	0.445	0.215	118.486	25.414
E+G+GE+BRR	Grain_Moisture_weight	MOH1_2016	3.665	5.633	128.039	0.651	0.044	0.029
E+G+GE+BRR	Grain_Moisture_weight	NCH1_2016	7.758	2.025	11.167	3.831	0.181	0.695
E+G+GE+BRR	Grain_Moisture_weight	NEH1_2016	113.669	56.705	50.862	2.005	1.115	2.235
E+G+GE+BRR	Grain_Moisture_weight	NYH2_2016	80.595	2.534	6.431	31.802	0.394	12.532
E+G+GE+BRR	Grain_Moisture_weight	OHH1_2016	12.108	4.124	14.744	2.936	0.280	0.821
E+G+GE+BRR	Grain_Moisture_weight	WIH1_2016	11.902	7.400	2.403	1.608	3.080	4.954
E+G+GE+BRR	Grain_Moisture_weight	WIH2_2016	0.917	4.113	0.764	0.223	5.385	1.201
E+G+GE+BRR	Grain_Moisture_weight	Across	–	–	–	5.085	7.900	3.409
E+G+GE+BRR	Yield_Mg_ha_BLUE	ARH1_2016	3.928	14.301	15.060	0.275	0.950	0.261

(Continued)

TABLE B6 Continued

Predictor	Trait	Env	NoEC	EC	FE	NoEC_vs_EC	EC_vs_FE	NoEC_vs_FE
E+G+GE+BRR	Yield_Mg_ha_BLUE	DEH1_2016	5.964	3.831	3.763	1.557	1.018	1.585
E+G+GE+BRR	Yield_Mg_ha_BLUE	GAH1_2016	3.379	4.157	4.699	0.813	0.885	0.719
E+G+GE+BRR	Yield_Mg_ha_BLUE	IAH1_2016	2.287	2.820	2.767	0.811	1.019	0.826
E+G+GE+BRR	Yield_Mg_ha_BLUE	IAH2_2016	7.505	7.733	8.012	0.971	0.965	0.937
E+G+GE+BRR	Yield_Mg_ha_BLUE	IAH3_2016	7.908	5.280	4.834	1.498	1.092	1.636
E+G+GE+BRR	Yield_Mg_ha_BLUE	IAH4_2016	2.565	3.811	3.842	0.673	0.992	0.668
E+G+GE+BRR	Yield_Mg_ha_BLUE	ILH1_2016	8.036	5.919	7.366	1.358	0.804	1.091
E+G+GE+BRR	Yield_Mg_ha_BLUE	INH1_2016	6.533	1.994	2.069	3.277	0.964	3.158
E+G+GE+BRR	Yield_Mg_ha_BLUE	MIH1_2016	4.748	19.667	20.508	0.241	0.959	0.232
E+G+GE+BRR	Yield_Mg_ha_BLUE	MNH1_2016	1.422	1.265	1.248	1.124	1.014	1.140
E+G+GE+BRR	Yield_Mg_ha_BLUE	MOH1_2016	12.381	9.392	11.632	1.318	0.807	1.064
E+G+GE+BRR	Yield_Mg_ha_BLUE	NCH1_2016	5.713	3.515	3.888	1.626	0.904	1.470
E+G+GE+BRR	Yield_Mg_ha_BLUE	NEH1_2016	5.446	5.214	4.593	1.045	1.135	1.186
E+G+GE+BRR	Yield_Mg_ha_BLUE	NYH2_2016	17.271	19.504	19.128	0.886	1.020	0.903
E+G+GE+BRR	Yield_Mg_ha_BLUE	OHH1_2016	2.503	2.138	2.234	1.171	0.957	1.121
E+G+GE+BRR	Yield_Mg_ha_BLUE	WIH1_2016	2.210	3.805	2.586	0.581	1.471	0.855
E+G+GE+BRR	Yield_Mg_ha_BLUE	WIH2_2016	4.667	5.288	5.442	0.883	0.972	0.858
E+G+GE+BRR	Yield_Mg_ha_BLUE	Across	–	–	–	1.117	0.996	1.095
E+G+GE+BRR	Yield_Mg_ha_weight	ARH1_2016	2.359	0.719	1.152	3.281	0.624	2.047
E+G+GE+BRR	Yield_Mg_ha_weight	DEH1_2016	0.051	0.020	0.186	2.540	0.108	0.273
E+G+GE+BRR	Yield_Mg_ha_weight	GAH1_2016	0.026	0.387	0.568	0.068	0.682	0.046
E+G+GE+BRR	Yield_Mg_ha_weight	IAH1_2016	2.914	2.808	2.836	1.038	0.990	1.027
E+G+GE+BRR	Yield_Mg_ha_weight	IAH2_2016	0.069	0.110	0.135	0.626	0.813	0.509
E+G+GE+BRR	Yield_Mg_ha_weight	IAH3_2016	0.670	1.666	3.383	0.402	0.493	0.198
E+G+GE+BRR	Yield_Mg_ha_weight	IAH4_2016	0.199	0.058	0.113	3.423	0.516	1.766
E+G+GE+BRR	Yield_Mg_ha_weight	ILH1_2016	0.751	0.594	0.808	1.264	0.736	0.930
E+G+GE+BRR	Yield_Mg_ha_weight	INH1_2016	0.112	0.099	0.064	1.130	1.550	1.750
E+G+GE+BRR	Yield_Mg_ha_weight	MIH1_2016	0.055	0.247	0.074	0.223	3.325	0.741
E+G+GE+BRR	Yield_Mg_ha_weight	MNH1_2016	0.146	0.338	0.673	0.432	0.502	0.217
E+G+GE+BRR	Yield_Mg_ha_weight	MOH1_2016	0.283	0.522	0.082	0.542	6.396	3.466
E+G+GE+BRR	Yield_Mg_ha_weight	NCH1_2016	0.113	0.227	0.100	0.499	2.277	1.135
E+G+GE+BRR	Yield_Mg_ha_weight	NEH1_2016	0.033	0.127	0.076	0.263	1.665	0.438
E+G+GE+BRR	Yield_Mg_ha_weight	NYH2_2016	0.449	0.418	0.553	1.074	0.756	0.812
E+G+GE+BRR	Yield_Mg_ha_weight	OHH1_2016	1.202	1.483	0.850	0.811	1.745	1.414
E+G+GE+BRR	Yield_Mg_ha_weight	WIH1_2016	0.204	0.110	0.066	1.858	1.672	3.107
E+G+GE+BRR	Yield_Mg_ha_weight	WIH2_2016	0.185	0.073	1.076	2.524	0.068	0.172
E+G+GE+BRR	Yield_Mg_ha_weight	Across	–	–	–	1.222	1.384	1.114

[4.045], MNH1_2016 [1.268], NYH2_2016 [16.252], OHH1_2016 [1.829] and WIH2_2016 [4.629]). On average, the RE values indicate general improvements of 10.650% for FE compared to EC, and 3.780% for EC compared to NoEC. However, when comparing the performance of NoEC and FE techniques, an average RE of 0.986 indicates a slight loss for FE compared to NoEC. For more detailed information, see [Table B6](#).

For the Yield_Mg_ha_weight trait, the use of NoEC achieved the best performance in most environments, as indicated by the MSE values (ARH1_2016 [0.989], GAH1_3016 [0.035], IAH2_2016 [0.175], IAH3_2016 [0.783], MIH1_2016 [0.1201], MHH1_2016 [0.393], MOH1_2016 [0.232], NYH2_2016 [0.402], OHH1_2016 [0.533] and WIH2_2016 [0.055]). On average, the RE values indicate a general improvement of 51.630% for EC compared to NoEC and 3.960% for FE compared to EC. However, when comparing the performance of NoEC and FE based on RE, the best performance was displayed by NoEC in most environments, resulting in an average RE of 0.9012, indicating that NoEC outperformed FE by 9.820%. For more detailed information, see [Table B6](#).

Predictor: E+G+GE+BRR

[Table B6](#) shows that EC yielded the most favorable results for the Grain_Moisture_BLUE trait in various environments. The corresponding MSE values for EC were 4.2801 (DEH1_2016), 0.519 (IAH4_2016), 4.964 (ILH1_2016), 4.762 (MNH1_2016), 6.047 (NEH1_2016), 4.030 (NYH2_2016), 3.072 (OHH1_2016), and 3.495 (WIH1_2016). Additionally, the average RE values indicated that using FE outperformed both EC and NoEC by 29.090% and 125.150%, respectively (1.291 for EC_vs_FE, and 2.252 for NoEC_vs_FE). Furthermore, when comparing the NoEC and EC techniques, an average RE of 1.6512 displays the superior performance of EC over NoEC by 65.180%. For more comprehensive information, see [Table B6](#).

When considering the Grain_Moisture_weight trait, the use of EC presented amor adequate performance in most environments

based on the MSE values provided in [Table 6](#) (DEH1_2016 [0.595], IAH1_2016 [360.363], IAH2_2016 [1.219], IAH4_2016 [120.354], ILH1_2016 [10.357], NCH1_2016 [2.0245], NYH2_2016 [2.534], and OHH1_2016 [4.124]). Moreover, the average RE values reveal that EC and FE outperformed the conventional NoEC by 408.510% and 240.900% respectively (5.085 for NoEC_vs_EC and 3.409 for NoEC_vs_FE). Furthermore, a comparison between EC and FE techniques indicates that an average RE of 7.899 suggests that FE outperformed EC by 689.960%. For more detailed information, see [Table B6](#).

When examining the Yield_Mg_ha_BLUE trait, the use of NoEC displayed the best performance in most environments based on the MSE values presented in [Table B6](#) (ARH1_2016 [3.928], GAH1_2016 [3.379], IAH1_2016 [2.287], IAH2_2016 [7.505], IAH4_2016 [2.565], MIH1_2016 [4.748], NYH2_2016 [17.271], WIH1_2016 [2.210] and WIH2_2016 [4.667]). However, it is worth noting that EC and FE outperformed the conventional NoEC by 11.690% and 9.490% in terms of average RE values (1.117 for NoEC_vs_EC and 1.095 for NoEC_vs_FE). Nevertheless, when comparing FE versus EC techniques, a slight loss of 0.400% was observed for using FE compared to EC, as indicated by an average RE of 0.996. For more detailed information, see [Table B6](#).

Regarding the Yield_Mg_ha_weight trait, [Table B6](#) shows that the use of EC yielded the best performance in most environments, as evidenced by the following MSE values: ARH1_2016 (0.719), DEH1_2016 (0.020), IAH1_2016 (2.808), IAH4_2016 (0.058), ILH1_2016 (0.594), NYH2_2016 (0.418), and WIH2_2016 (0.073). The average RE values indicated improvements of 22.200% (NoEC_vs_EC) and 11.380% (NoEC_vs_FE), highlighting the superior performance of EC and FE over the conventional NoEC technique. Conversely, when comparing EC and FE techniques, most environments performed better with an average RE of 1.384, indicating that FE outperformed EC by 38.420%. For additional information, see [Table B6](#).

TABLE B7 Variance components (Var_Comp) for environment (Env) Line and Genotype by environment (Env:Line) interaction for each data set. CV denotes coefficient of variation and n_Env denotes the average of number of environments in each data set.

Data	Component	VarComp	Trait	Heritability	CV	n_Env
Japonica	Env:Line	186065.908	GY	0.285	0.163	3.597
Japonica	Line	257287.998	GY	0.285	0.163	3.597
Japonica	Env	1860782.427	GY	0.285	0.163	3.597
Japonica	Residual	272836.420	GY	0.285	0.163	3.597
Japonica	Env:Line	0.000	PHR	0.462	0.073	3.597
Japonica	Line	0.000	PHR	0.462	0.073	3.597
Japonica	Env	0.001	PHR	0.462	0.073	3.597
Japonica	Residual	0.000	PHR	0.462	0.073	3.597
Japonica	Env:Line	0.000	GC	0.249	0.818	3.597
Japonica	Line	0.001	GC	0.249	0.818	3.597
Japonica	Env	0.006	GC	0.249	0.818	3.597

(Continued)

TABLE B7 Continued

Data	Component	VarComp	Trait	Heritability	CV	n_Env
Japonica	Residual	0.001	GC	0.249	0.818	3.597
Japonica	Env:Line	0.002	PH	0.624	0.097	3.597
Japonica	Line	20.528	PH	0.624	0.097	3.597
Japonica	Env	35.950	PH	0.624	0.097	3.597
Japonica	Residual	8.576	PH	0.624	0.097	3.597
USP	Env:Line	0.983	GY	0.533	0.378	4
USP	Line	1.129	GY	0.533	0.378	4
USP	Env	2.123	GY	0.533	0.378	4
USP	Residual	0.850	GY	0.533	0.378	4
G2F_2014	Env:Line	0.001	Grain_Moisture_BLUE	0.609	0.196	5.376
G2F_2014	Line	3.913	Grain_Moisture_BLUE	0.609	0.196	5.376
G2F_2014	Env	11.492	Grain_Moisture_BLUE	0.609	0.196	5.376
G2F_2014	Residual	2.006	Grain_Moisture_BLUE	0.609	0.196	5.376
G2F_2014	Env:Line	1.061	Grain_Moisture_weight	0.010	1.877	5.376
G2F_2014	Line	0.344	Grain_Moisture_weight	0.010	1.877	5.376
G2F_2014	Env	175.200	Grain_Moisture_weight	0.010	1.877	5.376
G2F_2014	Residual	3.331	Grain_Moisture_weight	0.010	1.877	5.376
G2F_2014	Env:Line	0.697	Yield_Mg_ha_BLUE	0.423	0.271	5.376
G2F_2014	Line	0.822	Yield_Mg_ha_BLUE	0.423	0.271	5.376
G2F_2014	Env	4.475	Yield_Mg_ha_BLUE	0.423	0.271	5.376
G2F_2014	Residual	0.853	Yield_Mg_ha_BLUE	0.423	0.271	5.376
G2F_2014	Env:Line	0.118	Yield_Mg_ha_weight	0.461	0.576	5.376
G2F_2014	Line	0.162	Yield_Mg_ha_weight	0.461	0.576	5.376
G2F_2014	Env	0.699	Yield_Mg_ha_weight	0.461	0.576	5.376
G2F_2014	Residual	0.202	Yield_Mg_ha_weight	0.461	0.576	5.376
G2F_2015	Env:Line	0.001	Grain_Moisture_BLUE	0.603	0.160	4.217
G2F_2015	Line	2.004	Grain_Moisture_BLUE	0.603	0.160	4.217
G2F_2015	Env	3.286	Grain_Moisture_BLUE	0.603	0.160	4.217
G2F_2015	Residual	2.270	Grain_Moisture_BLUE	0.603	0.160	4.217
G2F_2015	Env:Line	0.001	Grain_Moisture_weight	0.109	1.435	4.217
G2F_2015	Line	0.655	Grain_Moisture_weight	0.109	1.435	4.217
G2F_2015	Env	19.808	Grain_Moisture_weight	0.109	1.435	4.217
G2F_2015	Residual	2.699	Grain_Moisture_weight	0.109	1.435	4.217
G2F_2015	Env:Line	1.002	Yield_Mg_ha_BLUE	0.359	0.272	4.217
G2F_2015	Line	0.633	Yield_Mg_ha_BLUE	0.359	0.272	4.217
G2F_2015	Env	2.604	Yield_Mg_ha_BLUE	0.359	0.272	4.217
G2F_2015	Residual	1.164	Yield_Mg_ha_BLUE	0.359	0.272	4.217
G2F_2015	Env:Line	0.007	Yield_Mg_ha_weight	0.361	0.660	4.217
G2F_2015	Line	0.048	Yield_Mg_ha_weight	0.361	0.660	4.217

(Continued)

TABLE B7 Continued

Data	Component	VarComp	Trait	Heritability	CV	n_Env
G2F_2015	Env	0.284	Yield_Mg_ha_weight	0.361	0.660	4.217
G2F_2015	Residual	0.070	Yield_Mg_ha_weight	0.361	0.660	4.217
G2F_2016	Env:Line	0.000	Grain_Moisture_BLUE	0.830	0.142	10.055
G2F_2016	Line	2.387	Grain_Moisture_BLUE	0.830	0.142	10.055
G2F_2016	Env	3.584	Grain_Moisture_BLUE	0.830	0.142	10.055
G2F_2016	Residual	1.335	Grain_Moisture_BLUE	0.830	0.142	10.055
G2F_2016	Env:Line	0.014	Grain_Moisture_weight	0.109	1.259	10.055
G2F_2016	Line	0.468	Grain_Moisture_weight	0.109	1.259	10.055
G2F_2016	Env	34.317	Grain_Moisture_weight	0.109	1.259	10.055
G2F_2016	Residual	4.322	Grain_Moisture_weight	0.109	1.259	10.055
G2F_2016	Env:Line	1.477	Yield_Mg_ha_BLUE	0.736	0.252	10.055
G2F_2016	Line	1.337	Yield_Mg_ha_BLUE	0.736	0.252	10.055
G2F_2016	Env	2.211	Yield_Mg_ha_BLUE	0.736	0.252	10.055
G2F_2016	Residual	1.133	Yield_Mg_ha_BLUE	0.736	0.252	10.055
G2F_2016	Env:Line	0.020	Yield_Mg_ha_weight	0.341	0.598	10.055
G2F_2016	Line	0.023	Yield_Mg_ha_weight	0.341	0.598	10.055
G2F_2016	Env	0.372	Yield_Mg_ha_weight	0.341	0.598	10.055
G2F_2016	Residual	0.051	Yield_Mg_ha_weight	0.341	0.598	10.055



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Wenjun Huang,
Chinese Academy of Sciences (CAS), China
Deguo Han,
Northeast Agricultural University, China

*CORRESPONDENCE

Bing Hao

✉ Bing.Hao@hotmail.com

Shengchao Yang

✉ 13099437499@163.com

[†]These authors have contributed equally to
this work

RECEIVED 05 April 2024

ACCEPTED 20 May 2024

PUBLISHED 03 June 2024

CITATION

Song W, Zhang S, Li Q, Xiang G, Zhao Y,
Wei F, Zhang G, Yang S and Hao B (2024)
Genome-wide profiling of WRKY genes
involved in flavonoid biosynthesis in *Erigeron
breviscapus*.
Front. Plant Sci. 15:1412574.
doi: 10.3389/fpls.2024.1412574

COPYRIGHT

© 2024 Song, Zhang, Li, Xiang, Zhao, Wei,
Zhang, Yang and Hao. 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.

Genome-wide profiling of WRKY genes involved in flavonoid biosynthesis in *Erigeron breviscapus*

Wanling Song^{1,2†}, Shuangyan Zhang^{1,2†}, Qi Li^{1,2},
Guisheng Xiang^{1,2}, Yan Zhao^{1,2}, Fan Wei^{1,2},
Guanghui Zhang^{1,2}, Shengchao Yang^{1,2*} and Bing Hao^{1,2*}

¹The Key Laboratory of Medicinal Plant Biology of Yunnan Province, National & Local Joint
Engineering Research Center on Germplasm Innovation & Utilization of Chinese Medicinal Materials
in Southwest China, Yunnan Agricultural University, Kunming, China, ²Yunnan Characteristic Plant
Extraction Laboratory, Kunming, Yunnan, China

The transcription factors of WRKY genes play essential roles in plant growth, stress responses, and metabolite biosynthesis. *Erigeron breviscapus*, a traditional Chinese herb, is abundant in flavonoids and has been used for centuries to treat cardiovascular and cerebrovascular diseases. However, the WRKY transcription factors that regulate flavonoid biosynthesis in *E. breviscapus* remain unknown. In this study, a total of 75 *EbWRKY* transcription factors were predicted through comprehensive genome-wide characterization of *E. breviscapus* and the chromosomal localization of each *EbWRKY* gene was investigated. RNA sequencing revealed transient responses of 74 predicted *EbWRKY* genes to exogenous abscisic acid (ABA), salicylic acid (SA), and gibberellin 3 (GA3) after 4 h of treatment. In contrast, the expression of key structural genes involved in flavonoid biosynthesis increased after 4 h in GA3 treatment. However, the content of flavonoid metabolites in leaves significantly increased at 12 h. The qRT-PCR results showed that the expression patterns of *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* transcription factors exhibited a high degree of similarity to the 11 structural genes involved in flavonoid biosynthesis. Protein-DNA interactions were performed between the key genes involved in scutellarin biosynthesis and candidate WRKYs. The result showed that *F7GAT* interacts with *EbWRKY11*, *EbWRKY36*, and *EbWRKY44*, while *EbF6H* has a self-activation function. This study provides comprehensive information on the regulatory control network of flavonoid accumulation mechanisms, offering valuable insights for breeding *E. breviscapus* varieties with enhanced scutellarin content.

KEYWORDS

flavonoid, *Erigeron breviscapus*, WRKY, hormones, structure gene

1 Introduction

Erigeron breviscapus is a traditional medicinal plant in the Asteraceae family and is mainly distributed in Southwest China. The sales revenue of traditional Chinese medicine preparations derived from *E. breviscapus* as a primary ingredient in China reached 3 billion RMB in 2020 (<http://yn.chinadaily.com.cn/a/202007/27/WS5f1e9bb2a310a859d09da571.html>). The main active flavonoid in *E. breviscapus*, scutellarin, is extracted from the leaves and has been extensively utilized in prescription injections for the treatment of cardiovascular diseases (Chen et al., 2021; Ju et al., 2021; Yang et al., 2022b; Zhang et al., 2022). We successfully elucidated the complete biosynthesis pathway of scutellarin and constructed the high-level production yeast factory (Liu et al., 2018; Wang et al., 2022). Recently, we reported the transcription factors that regulate scutellarin (R2R3-MYB) and anthocyanin (bHLH) biosynthesis in *E. breviscapus* as well (Gao et al., 2022; Zhao et al., 2022). However, the regulatory mechanism of the flavonoid pathway governed by the WRKY transcription factor family in *E. breviscapus* remains elusive.

WRKY transcription factor is the seventh largest TF family in higher plants and is named for its characteristic WRKY domain (Rushton et al., 2010). The typical structure of WRKY is the N-terminal, which contains conserved amino acid sequence WRKYGQK, whereas the C-terminal contains a zinc finger motif (C2H2 or C2HC) (Eulgem et al., 2000). According to the number of WRKY domains and the type of zinc finger motif, WRKY can be divided into three categories: Group I contains two WRKY domains, whereas Groups II and III have a single WRKY domain, WRKY domains of Group II and III family members are more similar in sequence to the C-terminal than to the N-terminal WRKY domain of Group I proteins (Eulgem et al., 2000; Dong et al., 2003). The members of Group II WRKY were further divided into five subgroups: IIa, including IIa, IIc, IId, and IIe, based on additional conserved structural motifs (Eulgem et al., 2000; Zhang and Wang, 2005). WRKY transcription factors play important roles in plant growth and development, defense regulation, stress, and synthesis of secondary metabolites (Eulgem et al., 1999; Johnson et al., 2002; Yu et al., 2012; Yu et al., 2013).

Since the first WRKY gene (*SPL1*) was cloned from sweet potatoes (Ishiguro and Nakamura, 1994), the identification and functional analysis of WRKY genes has developed rapidly in plants, especially in crops, fruits, and medicinal plants. Several WRKY transcription factors have been identified in *Arabidopsis thaliana*, *Glycine max*, *Vitis vinifera*, *Panax ginseng*, and *Salvia miltiorrhiza*

(Wang et al., 2011; Guo et al., 2014; Li et al., 2015; Yang et al., 2017; Di et al., 2021). A total of 14549 WRKY genes were recorded in the Plant Transcription Factor Database (PlantTFDB) (Jin et al., 2017). Numerous studies have substantiated the close association between WRKY transcription factors and the biosynthesis of flavonoid metabolites (Amato et al., 2017; Duan et al., 2018; Wang et al., 2023). The overexpression of *AeWRKY32* (*Okra*) induced anthocyanin accumulation, with higher expression levels of *AtCHS1*, *AtCHI4*, *AtF3H1*, and *AtDFR2* in transgenic *Arabidopsis* (Zhu et al., 2023). *AtWRKY23* transcription factors regulate flavonol accumulation, auxin transport, root growth, and development (Grunewald et al., 2012). *VvWRKY70* and *NtWRKY11b* have been identified as regulators involved in flavonol biosynthesis, the content of flavonol significantly decreased in *VvWRKY70*-overexpressing grape calli lines by inhibiting the promoter *VvCHS2*, *VvCHS3*, and *VvFLS4*. Conversely, overexpression of *NtWRKY11b* led to a substantial increase in flavonol content ranging from 37.8% to 80.7%. (Wang et al., 2021; Wei et al., 2023). *BcWRKY1* significantly increased the transcript of *CHS* to regulate flavonoid biosynthesis (Zeng et al., 2022). However, the current literature rarely reports on the regulatory mechanisms by which WRKY transcription factors regulate flavone and flavonol biosynthesis in medicinal plants.

Plant hormones have a prominent function in the modulation of the growth, development, reproduction, and secondary metabolism of plants, such as SA, ABA, GA3, and MeJA, shown to be involved in the regulation of flavonoid biosynthesis (Khan et al., 2015; Lucho-Constantino et al., 2017; Li and Ahammed, 2023). Exogenous ABA could promote the synthesis of ABA in *Artemisia argyi* leaves and up-regulated the content of chlorogenic acid, nevertheless significantly down-regulated other flavonoid metabolites after ABA treatment (Yang et al., 2022a). Exogenous GA3 evidently decreased the contents of naringin and naringenin in *P. chinense* Schneid seedlings (Yang et al., 2023). Recently, an increasing interest has focused on WRKY transcription factors response to plant hormones and involved in flavonoid metabolism (Schlutenhofer and Yuan, 2015; Vives-Peris et al., 2018; Xu et al., 2020; Yamamoto et al., 2020). *LrWRKY3* transcription factor response to MeJA may specifically interact with the *ANR* and *LAR* gene and might be involved in anthocyanins synthesis in *L. radiata*, regulated the content of pelargonidin-3-O-glucoside-5-O-arabinoside in *L. radiata* (Wang et al., 2023). *VqWRKY31* also activated SA defense signaling and changed the accumulation of stilbenes, flavonoids, and proanthocyanidins (Yin et al., 2022). Flavonoid compounds are involved in the defence of plants against biotic and abiotic stresses, WRKY transcription factors can respond to hormone signal transduction pathways, improve the accumulation of flavonoids, and play key roles in the regulation of various stressful stresses (drought, low temperature, wounds, disease-resistant, etc.) in plants (Pourcel et al., 2007; Agati et al., 2012; Han et al., 2018a, Han et al., 2018b). The main medicinal active ingredients of *E. breviscapu* are flavonoids, the study of the response mechanisms of WRKY transcription factors and flavonoids under hormonal stress, and can effectively analyze and identify WRKY transcription factors involved in the synthesis of scutellarin, and elucidate the molecular

Abbreviations: ABA, abscisic acid; SA, salicylic acid; GA3, gibberellin 3; *AtWRKY*, *Arabidopsis thaliana* WRKY Transcription factor; *EbWRKY*, *Erigeron breviscapus* (Vant.) Hand.-Mazz WRKY Transcription factor; *HaWRKY*, *Helianthus annuus* WRKY Transcription factor; qRT-PCR, Quantitative RT-PCR; *PAL*: Phenylalanine ammonia-lyase; *C4H*: Cinnamate 4-hydroxylase; *4CL*, 4-coumarate, CoA ligase; *CHS*, Chalcone synthase; *CHI*, Chalcone isomerase; *FSII*, Flavone synthase II; *EbF6H*, Flavone 6-hydroase; *F7GAT*: Flavonoid 7-O-glucuronosyltransferase; *F3H*: Flavanone 3-hydroxylase; *F3'H*, Flavonoid 3'-hydroxylase; *FLS*, Flavonol synthase; UPLC, Ultra-high-performance liquid chromatography; SE, Scutellarin.

mechanisms by which WRKY transcription factors regulate the synthesis of scutellarin.

In this study, the conserved motifs, gene structure, chromosome location, phylogenetic trees, gene expression profile, and function of WRKY genes were identified based on the whole genome of *E. breviscapus*. Additionally, integrated metabolomic and transcriptomic analyses were performed to study the expression patterns of WRKY genes and flavonoid metabolites in response to exogenous hormone treatments. Our study revealed the expression of *EbWRKYs* and the accumulation of flavonoids differed under the treatment of three exogenous hormones in *E. breviscapus* leaves. We also identified three candidate WRKY genes potentially involved in the regulation of scutellarin biosynthesis. This study provided valuable guiding information for growth and development research and functional identification of WRKY transcription factors involved in scutellarin biosynthesis in *E. breviscapus*.

2 Methods

2.1 Plant treatment

The two-month-old *E. breviscapus* seedlings were germinated and cultivated in a growth chamber under controlled conditions at 22 °C with a photoperiod of 16 hours light and 8 hours dark. Furthermore, the leaves of *E. breviscapus* were treated with 200 mL of ABA, SA, and gibberellin 3 (GA3) solution at a concentration of 200 µmol/L. Leaf samples were collected at time points of 0 h, 4 h, 12 h, and 24 h, immediately frozen in liquid nitrogen, and stored at -80 °C. Each experimental sample has three biological repetitions.

2.2 Identification and physicochemical properties of WRKY proteins

The PfamScan v1.6 tool was employed to annotate the protein domains of the entire genome sequence of *E. breviscapus*, utilizing the Pfam 35.0 database. Sequences exhibiting E-values lower than 10^{-5} and encompassing the PF03106 domain were screened, while manually excluding any atypical characteristics observed in WRKY genes. The ProtParam tool (<https://web.expasy.org/protparam/>) was utilized for predicting various attributes of *EbWRKY* proteins, including molecular weights (MWs), isoelectric points (pIs), amino acid counts, open reading frame (ORF) lengths. Protein subcellular localization was predicted by PSORT (<https://psort.hgc.jp/>).

2.3 Protein domain and phylogenetic evolution analysis

Multiple sequence alignments were conducted using MAFFT v7.490 to elucidate the evolutionary relationship between *E. breviscapus* and *A. thaliana*. To investigate the interrelationship among *E. breviscapus* WRKY proteins, a phylogenetic tree

encompassing both *E. breviscapus* and *A. thaliana* WRKY proteins was constructed through PhyIip v3.698 software employing the neighbor-joining method with 1000 repetitions. Subsequently, EvolView (<https://evolgenius.info/evolview/#/>) was employed as an evolutionary tree visualization tool for further analysis. The WRKY protein sequences of *A. thaliana* were downloaded from the TAIR database (<https://www.arabidopsis.org/>).

2.4 Comprehensive analysis of WRKY genes

The intron-exon structures of the *EbWRKY* genes were determined using the gene structure display server provided by Peking University (<http://gsds.cbi.pku.edu.cn/>). The conserved motifs of WRKY proteins were predicted using Motif Elicitation, a tool available at <http://alternate.meme-suite.org/tools/meme>. For motif prediction, we employed optimized parameters including any number of repetitions (20), minimum width (10), and maximum width (80). Gene structure and chromosome mapping analysis of WRKY family members were conducted using TBtools v1.098691, while collinearity between the WRKY gene family in *E. breviscapus* and *A. thaliana*, *Daucus carota*, *Helianthus annuus*, *Lycopersicon esculentum*, and *Solanum tuberosum* was analyzed with the one-step MCScanX tool. Collinearity within *E. breviscapus* was visualized using Advanced Circos software.

2.5 RNA-sequencing data analysis

The leaves of *E. breviscapus* were subjected to hormone treatment, followed by flash freezing in liquid nitrogen for RNA extraction and subsequent cDNA library construction. Transcriptome sequencing was performed using Illumina HiSeq 4000 platform. SkrTools (version 1.0) was employed to calculate the raw data generated from sequencing, which underwent filtration using Trimmomatic v0.39 and RiboDetector v0.2.4 (Bolger et al., 2014; Deng et al., 2022). Subsequently, rRNA sequences were eliminated from the raw data to obtain high-quality clean reads that were utilized for gene differential expression analysis.

2.6 Expression profiling analysis of *EbWRKY* genes in various tissues

The differential expression of *EbWRKYs* in roots, stems, leaves, and flowers was calculated using the Salmon software based on the previously assembled genomic data of *E. breviscapus*. Additionally, the expression levels in the transcriptome data treated with three exogenous hormones at different time points were analyzed. The TPM value (Transcripts Per Million) was used to calculate gene expression values. Clustering results and heat maps were generated using TBtools v1.098691 (Chen et al., 2020).

2.7 Metabolites analysis

The frozen and fresh leaves of *E. breviscapus* (100 mg) were ground in liquid nitrogen, and the homogenate was resuspended in pre-chilled 80% methanol and 0.1% formic acid by vortexing. The samples were incubated on ice for 5 min and then centrifuged at 15,000 g at 4°C for 20 min. The supernatant was diluted to a final methanol concentration of 53% for the LC-MS/MS analysis (Dunn et al., 2011). Samples were injected onto an Xselect HSS T3 column (2.1×150 mm, 2.5 μm) with a 20-min linear gradient at a 0.4 mL/min flow rate for the positive/negative polarity mode. The eluents used were eluent A (0.1% formic acid water) and eluent B (0.1% formic acid-acetonitrile). The solvent gradient was set as follows: 2% B, 2 min; 2–100% B, 15.0 min; 100% B, 17.0 min; 100–2% B, 17.1 min; 2% B, 20 min (Wang et al., 2014). The data files generated by HPLC-MS/MS were processed using the SCIEX OS Version.

The dried leaves of *E. breviscapus* powder sample (0.3 g) were dissolved in 50 mL of methanol, and the supernatant was extracted for 30 minutes by ultrasonic use for HPLC analysis. Samples were injected onto an Agilent EC-C18 column (4.6 x 100 mm, 2.7 μm), with a 50-min linear gradient at a 1 mL/min flow rate. The eluents used were eluent A (acetonitrile) and eluent B (0.1% phosphoric acid water). The solvent gradient was set as follows: 0–10 min, 12%–15% A; 10–32 min, 15% A; 32–33 min, 15%–20% A; 33–50 min, 20%–22% A. Scutellarin (SE), Chlorogenic acid (CGA), 3,5-dicaffeoylquinolinic acid (3,5-diCQA), and Erigeron B (EB) were quantified using the external standard method with standards purchased from Sigma-Aldrich (Shanghai, China). Variance significance analysis was conducted employing SPSS 20.0.

2.8 Co-expression network analysis

The expression of candidate *EbWRKYs* and key genes involved in flavonoid biosynthesis was extracted from the transcriptome data obtained from the roots, stems, leaves, and flowers of *E. breviscapus*. Initially, statistically significant correlations between differential metabolites were calculated using R (version 4.1.1). A significance level of $p < 0.05$ was applied for statistical analysis. Subsequently, gene expression levels and relative metabolite contents were collected to identify correlation pairs with a Pearson product-moment correlation coefficient (PCC) ≥ 0.6 and a p -value ≤ 0.05 . The filtered genes were then utilized to construct a co-expression network which was visualized using Cytoscape version 3.3.0 software (<https://www.cytoscape.org>).

2.9 Protein-DNA interactions assays

The proteins-DNA interaction between *EbWRKY11*, *EbWRKY36*, *EbWRKY44*, *EbF6H*, and *F7GAT* was investigated using the bait construct pAbAi and prey construct PGADT7, which were generated through a BP reaction. The prey plasmid was transformed into the bait strain yeast Y1H and selected with supplemented medium containing SD/-Leu, SD/-Leu/Aba

(Clontech). The binding domain was predicted using JASPAR (<https://jaspar.elixir.no/>).

2.10 Quantitative real-time PCR analysis

Leaves were collected, and total RNA was extracted from hormone-treated samples using a HiPure HP Plant RNA Mini Kit (R4165-02). Subsequently, cDNA synthesis was performed utilizing a PrimeScript RT Reagent Kit with gDNA Eraser (Takara, Japan). Gene-specific primers for qRT-PCR reactions were designed employing Primer3 web version 4.1.0 (<https://primer3.ut.ee/>) (Supplementary Table S6). A Quantstudio 5 Flex Real-Time PCR System (Thermo Fisher Scientific, USA) was employed to analyze three technical replicates. The expression levels of genes from different treatments were normalized to *EbACTIN2*. Finally, the relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method and visualized using GraphPad Prism 8.0.2.

3 Results

3.1 Identification and physicochemical properties of WRKY genes in *E. breviscapus*

A total of 75 putative *EbWRKYs* were identified from *E. breviscapus* genomic data, which were designated as *EbWRKY1* to *EbWRKY75*. The number of amino acids ranged from 144 (*EbWRKY51*) to 756 (*EbWRKY19*), and the isoelectric points ranged from 4.97 (*EbWRKY29/47*) to 10.12 (*EbWRKY73*), including 45 acidic and 30 basic amino acids. In addition, the relative molecular weights ranged from 19.99 (*EbWRKY68*) to 80.89 kDa (*EbWRKY25*). Based on sequence analysis conducted using PSORT software, it was determined that the 75 *EbWRKY* proteins are localized within the nucleus, suggesting their potential regulatory roles as transcription factors in this cellular compartment. In addition, *EbWRKY19* is located on the cell membrane and may be involved in the expression and regulation of genes related to membrane transport (Supplementary Table S1).

3.2 Evolution and sequence analysis of WRKY transcription factors

To gain a comprehensive understanding of plant biodiversity mechanisms and the regulatory role of WRKY genes in the network, we conducted an evolutionary analysis of WRKY transcription factors. Subsequently, phylogenetic analyses were performed on 75 *E. breviscapus* and 72 *Arabidopsis* WRKY transcription factors (Figure 1A). A total of 147 WRKYs were divided into seven branches. *E. breviscapus* and *Arabidopsis* WRKY proteins with the same classifications were classified into I, II, and III. Group I has 17 *EbWRKYs*, Group II can be divided into five subtypes according to the different zinc-finger structural sites: IIa, IIb, IIc, IId, and IIe. There were five *EbWRKYs* in Group IIa, eight *EbWRKYs* in Group IIb, eleven *EbWRKYs* in Group IIc, ten *EbWRKYs* in Group IId,

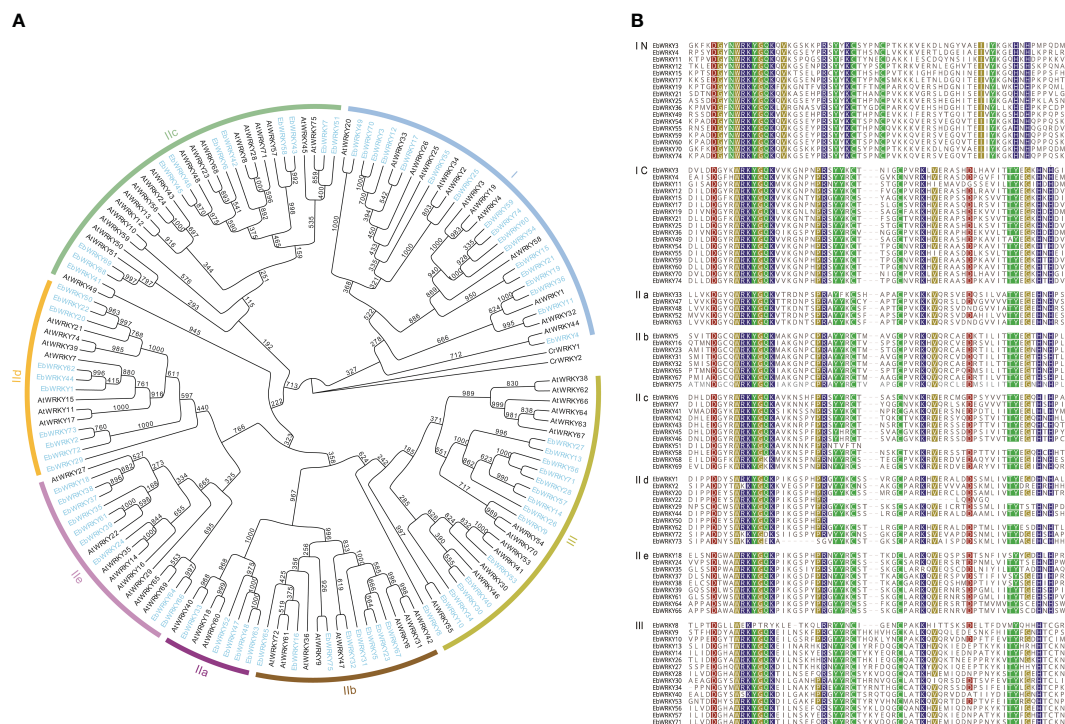


FIGURE 1

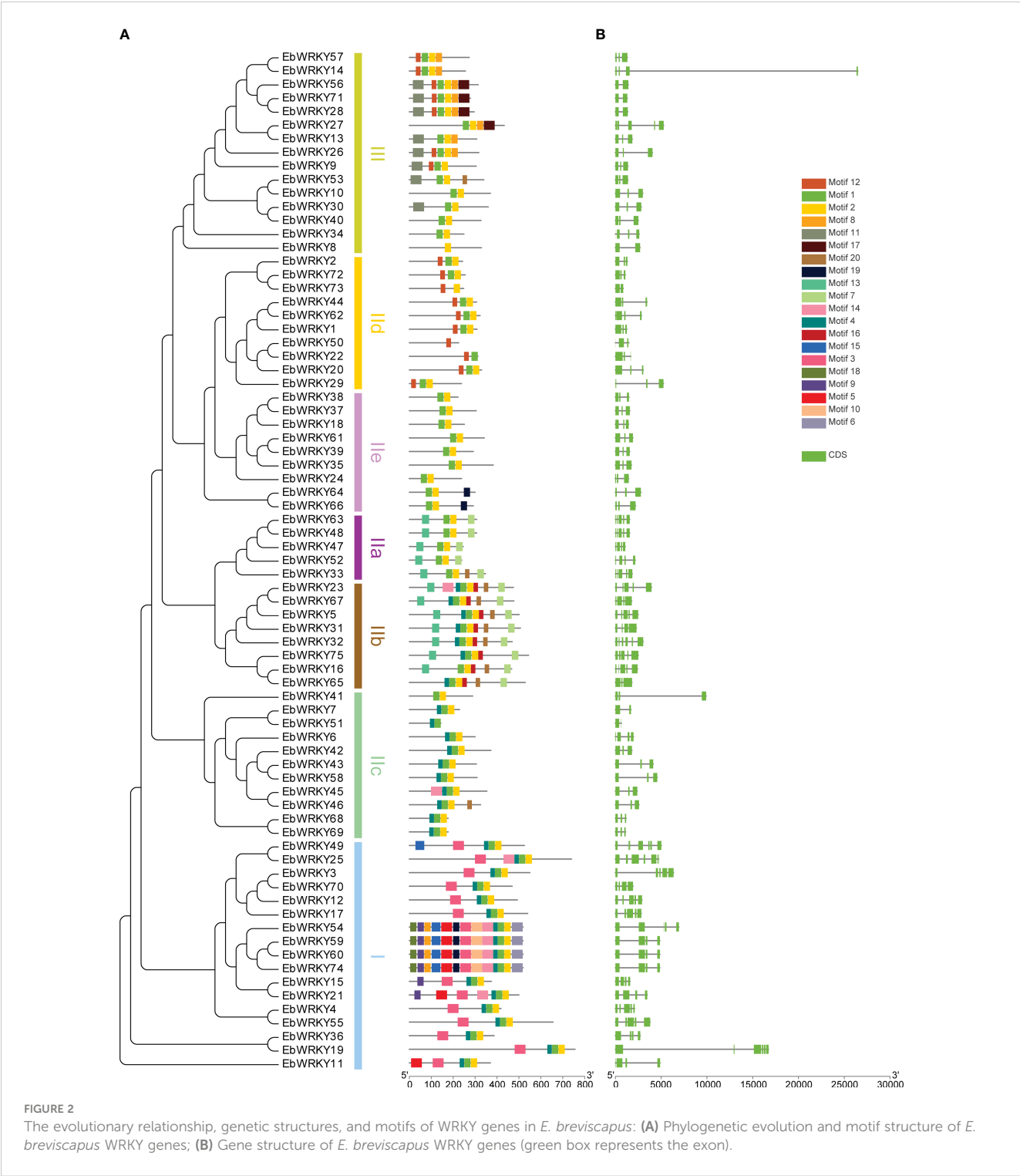
Phylogenetic and WRKY protein domain sequence analysis of *E. breviscapus*: (A) Phylogenetic analysis of *E. breviscapus* and *Arabidopsis* WRKY transcription factor; (B) WRKY protein domain sequence analysis in *E. breviscapus*. IN and IC represent the C-terminal and N-terminal WRKY domain of Group I, respectively.

nine *EbWRKYs* in Group IIe, and fifteen *EbWRKYs* in Group III. The three major classes and five subclasses in the phylogenetic tree contained both the WRKY genes from *E. breviscapus* and *Arabidopsis*, indicating that the WRKY families of *Arabidopsis* and *E. breviscapus* are highly similar at the evolutionary level. Additionally, the WRKY transcription factors of *E. breviscapus* exhibit a high degree of similarity within their respective branches of the phylogeny, suggesting an increased homogeneity in the WRKY gene family during evolutionary processes.

In addition to WRKYGQK, the core motif of *E. breviscapus* WRKY heptapeptide contained five variants: WRKYGKK, WKKYGQK, WKKYGDK, WKKYGEK, and WSKYGQK (Figure 1B). Sequence comparison results showed that each WRKY protein, except *EbWRKY8*, contained a typical WRKY conserved domain at the N-terminal and a complete zinc finger structure at the C-terminal (CX4-5CX22-23HXH/C), which is an important feature for identifying WRKY transcription factors. All WRKY sequences of *E. breviscapus* showed a high similarity and conservation of the WRKY domain. Group I *EbWRKYs* contained the same heptapeptide core motif WRKYGQK at the N- and C-terminal and the zinc finger structure C2H2 behind the WRKY structure at the C-terminal. Group II and Group III had a WRKY domain at the N-terminal, but the zinc finger structure at the C-terminal differed (C2HX). *EbWRKY8* was domain sequence was lost, and the C-terminal retained a zinc-finger structure. However, the sequences of *EbWRKY8* are highly similar to the other *EbWRKYs*, and evolutionary analysis clustered them into Group III.

3.3 Gene structure and conserved motif analysis of WRKY proteins

75 WRKY protein motifs were analyzed using MEME and TBtools. The conserved domain of the WRKY motif was identified in motifs 1 and 3, while motif 2 exhibited a zinc-finger structure. Motifs 1 and 2 were found in almost all *EbWRKY* proteins, indicating their widespread presence. Notably, distinct *EbWRKYs* displayed diverse motif structures, with similar motifs observed within each branch clustering by the same type of *EbWRKY*. Motif 15 was exclusively present in the transcription factor genes of *EbWRKY49*, *EbWRKY54*, *EbWRKY59*, *EbWRKY60*, and *EbWRKY74* in Group I. Motif 16 was identified as a characteristic motif of Group IIb *EbWRKYs* while motif 4 was found to be a common feature of Groups I and II (b,c). Only eight members of Group III contained motif 11 whereas motif 12 was detected both in the sequences of Groups III and IId *EbWRKY* (Figure 2A; Supplementary Table S2). Phylogenetic trees of WRKY proteins were established, and the three groups were clustered according to their sequence similarity. WRKY sequences with similar structures in the evolutionary tree clustered into a single branch, indicating that these WRKY proteins may have similar functions. TBtools were used to analyze the number and distribution of exons of CDS sequences of the 75 WRKYs (Supplementary Table S3). The results showed a significant difference in the number of introns and exons in the WRKY gene family of *E. breviscapus*. The numbers of introns and exons were 1-6 and 2-7. (Figure 2B).



3.4 Chromosomal mapping and collinearity analysis of WRKY genes

The distribution of the 75 *EbWRKY* genes across all nine *E. breviscapus* chromosomes exhibited irregular patterns (Figure 3A). A total of 21 *EbWRKY* genes were localized to chromosome 1, accounting for 28% of the *EbWRKY* gene family. Ten tandem duplications occurred on the six chromosomes. Eight WRKY

members were collinear on chromosomes 1, 3, 4, and 6, respectively (Figure 3B). To further predict the potential evolutionary patterns of the *EbWRKY* gene family, we constructed comparative syntenic maps of *E. breviscapus* in associated with five representative species, including *A. thaliana*, *D. carota*, *H. annuus*, *L. esculentum*, and *S. tuberosum* (Figure 3C). The number of orthologous gene pairs between *E. breviscapus* and *A. thaliana*, *D. carota*, *H. annuus*, *L. esculentum*, and *S. tuberosum*

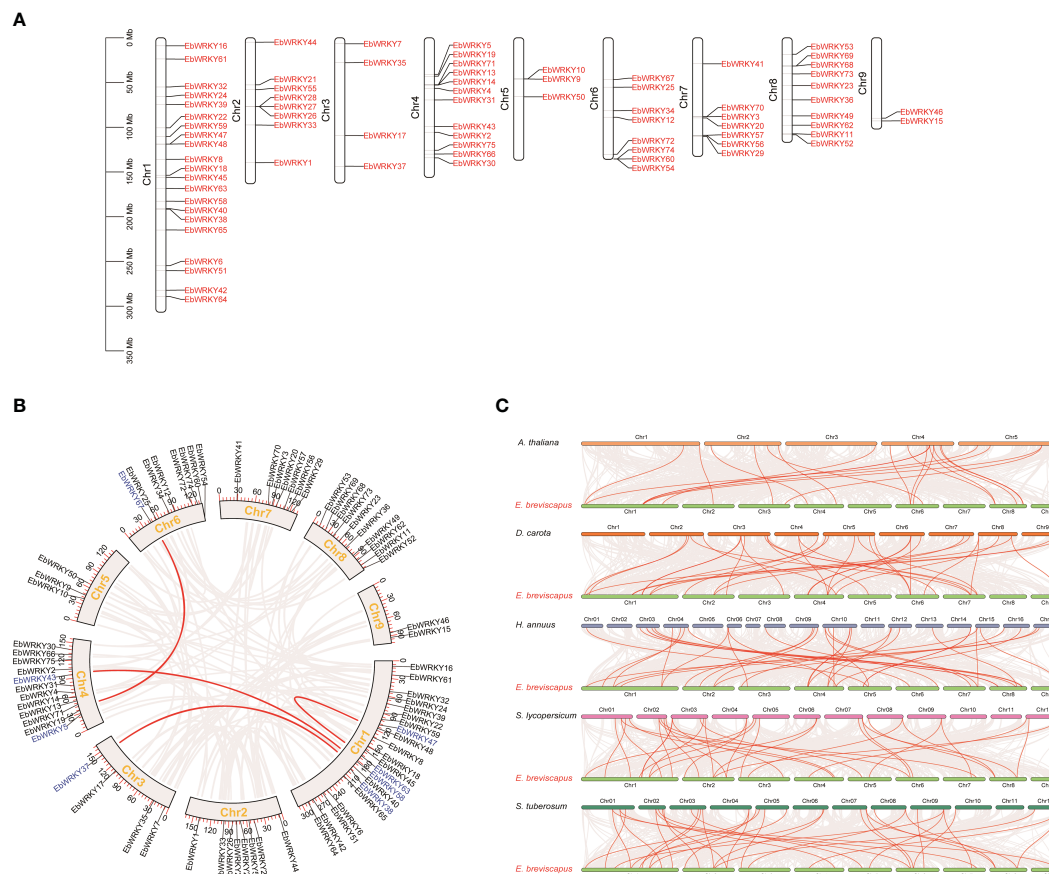


FIGURE 3

The Synteny analysis and chromosome location of WRKY genes in *E. breviscapus*: (A) Chromosomal location of WRKY genes; (B) The internal collinearity circle diagram of the *E. breviscapus* genome (the black line represents the position of genes on chromosomes; the arc represents the collinearity relationship between genes; the WRKY gene pair is highlighted with a red line; the gene name color represents different subgroups); (C) WRKY gene pairs are highlighted with red lines according to the analysis of collinearity between *E. breviscapus* and different plant genomes.

were 21, 37, 33, 34, and 36, respectively (Supplementary Table S4). These results revealed that the identified orthologous events of *EbWRKY-HaWRKY* were considerably higher than those of other WRKY species based on their close evolutionary relationship. An extensive level of synteny conservation and an increased number of orthologous events in *EbWRKY-HaWRKY* indicated that *EbWRKY* genes in *E. breviscapus* shared a similar structure and function with *HaWRKY* genes.

3.5 Expression pattern of WRKY genes in the different tissues and hormone treatment

The TPM values of the roots, stems, leaves, and flowers were extracted from the genomic database to clarify the expression of *EbWRKY* family genes. Except for seven *EbWRKY* genes that exhibited no expression in any tissue, the remaining *EbWRKY* genes demonstrated specific expression patterns in leaves, roots,

stems, and flowers, respectively. (Supplementary Table S5; Figure 4A). In our previous study, exogenous application of SA, GA3, and ABA onto the leaves of *E. breviscapus* was found to enhance their scutellarin (SE) content, with ABA treatment showing the most significant effect (Supplementary Table S6). Therefore, transcriptomic analysis was employed to investigate the underlying expression mechanism of *EbWRKYs* in response to hormone induction. The results showed that the expression levels of *EbWRKYs* significantly altered after three hormone treatments. In the ABA treatment, the expression levels of *EbWRKY69* and *EbWRKY30* were significantly up-regulated after 4 h. In contrast, in the SA treatment assays, the gene expression levels of *EbWRKY8*, *EbWRKY17*, *EbWRKY51*, *EbWRKY67*, *EbWRKY18*, *EbWRKY66*, and *EbWRKY64* were significantly up-regulated after 12 h of treatment, and *EbWRKY52* and *EbWRKY57* were significantly up-regulated after 24 h. The expression levels of *EbWRKY3*, *EbWRKY41*, *EbWRKY47*, *EbWRKY2*, and *EbWRKY39* genes were significantly upregulated after 4 hours of GA treatment. However, a gradual decline in their expression levels was observed in the leaves over time. (Figure 4B).

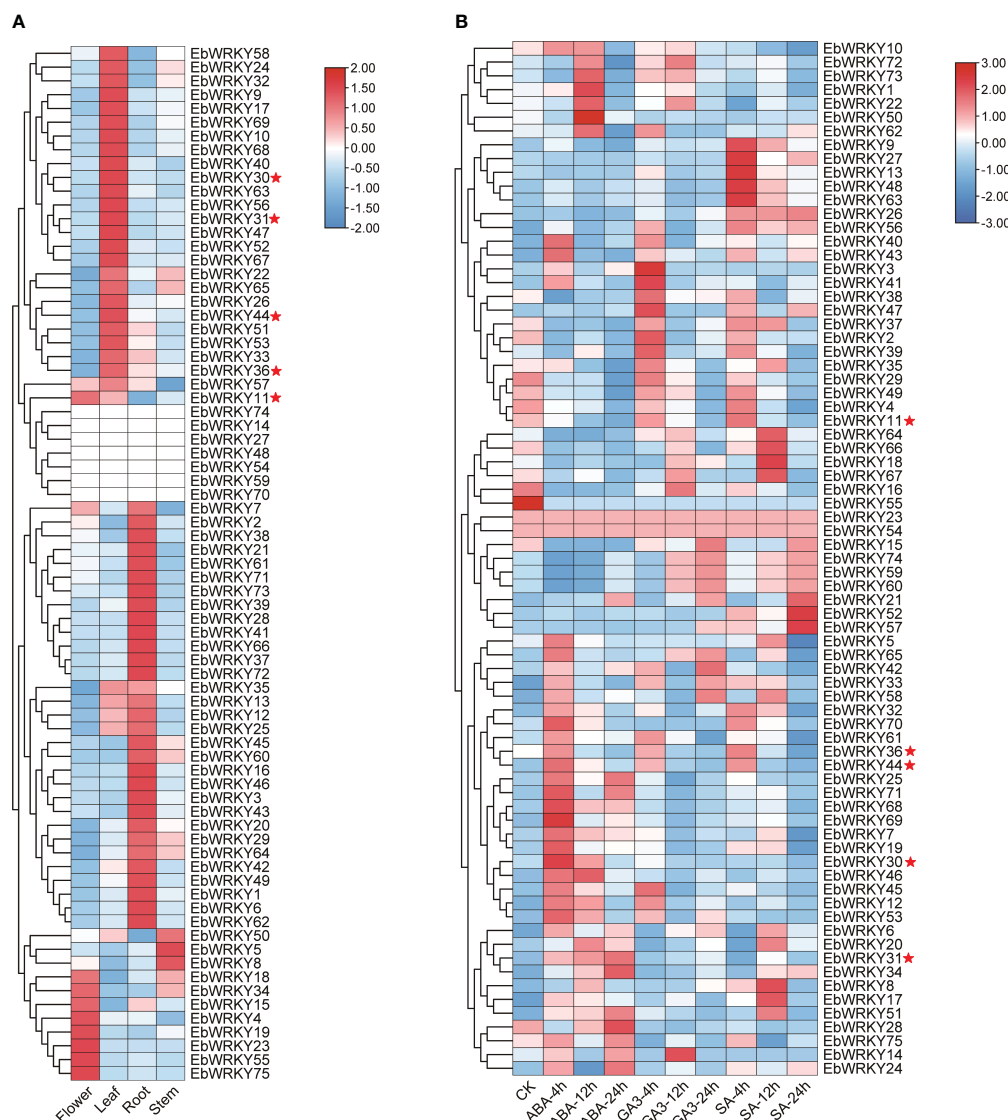


FIGURE 4

The expression profiles of WRKY genes in *E. breviscapus*: (A) Expression profiles of WRKY genes in different tissues of *E. breviscapus*; (B) Expression profiles of *E. breviscapus* WRKY under different hormone treatments at different times. The color scale on the right of each diagram represents TPM expression values: red indicates higher levels and blue indicates lower levels. CK represents the untreated sample. The red star represents five candidate *EbWRKYs* that may be involved in the flavonoid metabolic pathway in *E. breviscapus*.

3.6 Hormone-induced expression analysis of structural gene and flavonoid metabolites in the leaves *E. breviscapus*

Ultra-high-performance liquid chromatography (UPLC) and tandem mass spectrometry (MS/MS) were used to determine dynamic changes in flavonoid metabolites in nine *E. breviscapus* treated with the three hormones. Pearson correlation coefficients of the QC samples were calculated based on the relative quantitative values of the metabolites. The R^2 values of all the samples were close to 1, indicating better stability of the entire detection process and higher data quality (Supplementary Table S7; Figure 5A). Scutellarin biosynthesis commences with phenylalanine, followed by the enzymatic catalyzation of phenylalanine ammonia-lyase (PAL), cinnamate 4-hydroxylase (C4H), coumaric acid coenzyme

A ligase (4CL), chalcone synthase (CHS), chalcone isomerase (CHI), flavone synthase II (FSII), flavonoid 7-O-glucuronosyltransferase (F7GAT), and flavone-6 hydroxylase (F6H) to yield scutellarin. Furthermore, other flavonoids including kaempferol, quercetin hesperidin, and luteolin are biosynthesized via the flavanone 3-hydroxylase (F3H), flavonoid 3'-hydroxylase (F3'H), and flavonol synthase (FLS).

The expression patterns of 11 key enzyme genes involved in flavonoid biosynthesis were analyzed following treatments with ABA, SA, and GA3. The results showed that the expression of 11 genes treated with the three hormones was up-regulated at 4 h. The downstream genes regulating flavonol biosynthesis, *F3H* and *F3'H*, were significantly up-regulated in ABA treatment at 24 h but down-regulated in GA3 treatment. The gene expression of GA3 exhibited its peak at 4 hours, while the response to SA did not

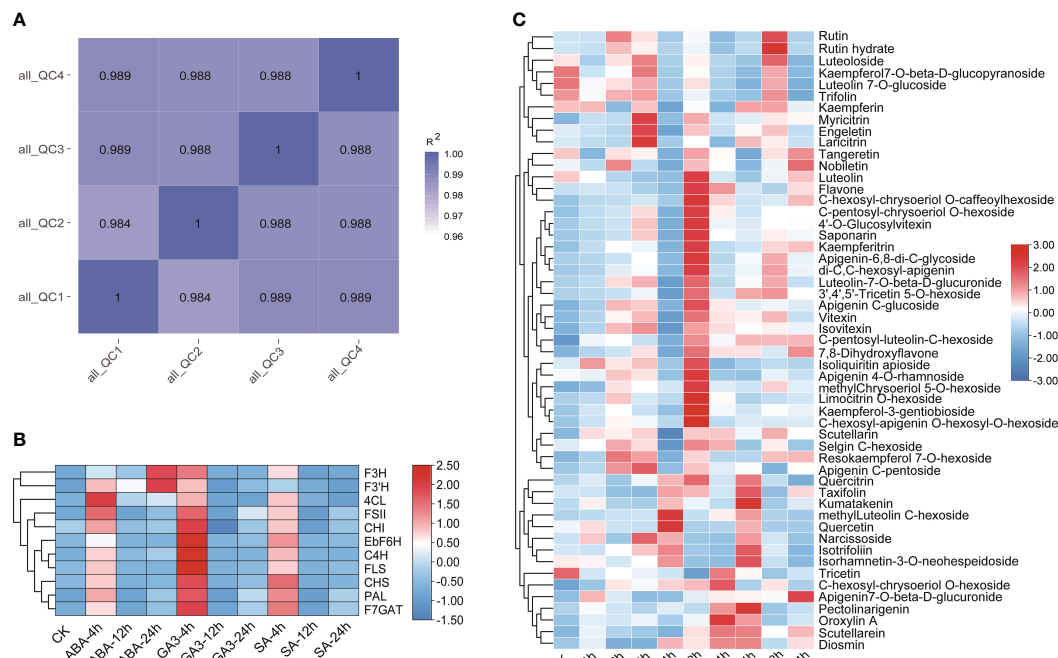


FIGURE 5

Quantitative and structural gene expression analysis of flavonoid pathway metabolites in the leaves of *E. breviscapus*: (A) Pearson correlation coefficient analysis of the QC of samples; (B) Structural gene expression analysis; (C) Contents of 54 flavone and flavonol metabolites in *E. breviscapus* under abscisic acid (ABA), salicylic acid (SA), and gibberellin 3 (GA3) hormone treatments at different times. CK represents the unprocessed sample.

manifest prominently. The expression levels of *FLS*, *C4H*, and *F6H* were significantly up-regulated after 4 h of GA3 treatment (Supplementary Table S8; Figure 5B).

A total of 159 flavonoids were identified, with three biological replicates set for each sample (Supplementary Table S7). Flavones and flavonols accounted for the majority (57.8%), followed by flavanones (15.7%), isoflavones (10.6%), and anthocyanins (8.1%). Chalcones, dihydrochalcones (4.4%), and other flavonoids (3.1%) were found in lower abundance. The analysis revealed the identification of 92 flavone and flavonol metabolites, with scutellarin exhibiting the highest content, followed by Apigenin7-O-β-D-glucuronide (Supplementary Table S7; Figure 5C). After treatment with ABA, SA, and GA3, the metabolism patterns of 54 flavone and flavonol metabolites exhibited differential changes, and the responses of flavonol and flavone compounds significantly increased after 12 h of GA3 treatment, with 14 compounds significantly increased compared to other levels. In addition, the content of pectolinarigenin significantly increased after 24 hours of SA treatment, while the content of oroxylin A showed a significant increase after 24 hours of GA3 treatment compared to other levels.

3.7 Integrated analysis of WRKYs involved in flavonoid metabolism

Transcriptome and metabolome data were integrated and analyzed to construct a co-expression network of key genes involved in flavonoid metabolism pathways. A co-expression

network was constructed by screening relevant pairs through Pearson analysis of gene expression levels and compound content ($PCC \geq 0.6$, $p \leq 0.05$). A total of 231 related pairs were identified and visualized using Cytoscape software (version 3.3.0). The network revealed a total of 102 interconnected nodes connected by 231 edges, encompassing 10 key enzyme genes and 45 *EbWRKYs*, alongside the presence of 47 flavonoid metabolites comprising 26 flavone and flavonol derivatives. In addition, a positive correlation was observed in 143 pairs, while 88 pairs exhibited a negative correlation (Figure 6). Within the flavonoid metabolic pathway, five potential *EbWRKYs* were identified as candidates with positive associations, namely *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44*. Notably, among these candidates, *EbWRKY11* demonstrated connections to four key genes (*EbF6H*, *F7GAT*, *FLS*, and *CHI*). The expression of *EbWRKY30* showed a positive correlation with *4CL*, *EbWRKY31*, and *F3'H*. Additionally, the presence of *EbWRKY36* and *EbWRKY44* was found to be associated with *PAL*, *EbF6H*, *C4H*, *4CL*, *F7GAT*, *CHI*, *CHS*, and *FLS*. Notably, *EbWRKY44* also exhibited a connection with *FSII* (Figure 6).

Additionally, *EbWRKY11* exhibited association with a total of 17 flavonoid metabolites, encompassing nine flavone and flavonol metabolites. Among the 13 *EbWRKYs*, eight pairs displayed positive correlation while five pairs showed negative correlation. *EbWRKY31* was associated with six metabolites, including three flavones and flavonols. Seven pairs were positively correlated, and eight pairs were negatively correlated with 15 *EbWRKYs*. However, *EbWRKY30* was not directly connected with metabolites but was

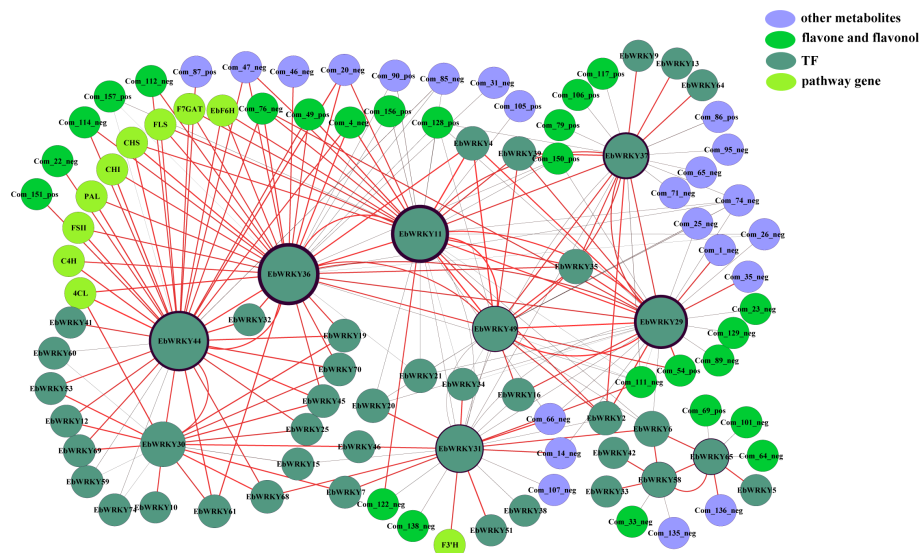


FIGURE 6

Co-expression analysis of structural genes involved in the flavonoid biosynthesis pathway and *EbWRKYs* in the leaves of *E. breviscapus*. Yellowish-green nodes represent genes; Blackish-green nodes represent WRKY TFs; Lavender nodes represent other flavonoid metabolites; Green nodes represent flavone and flavonol metabolites. The size of the circle is associated with the number of *EbWRKY* genes. Black circles outside the genes are associated with the number of metabolites.

related to 17 *EbWRKYs*. The *EbWRKY36* gene was found to be associated with 18 flavonoid metabolites and 12 other *EbWRKY* genes, exhibiting positive correlations in 11 pairs and a negative correlation in one pair. The correlation analysis revealed that *EbWRKY44* was associated with 12 flavonoid metabolites and sixteen *EbWRKYs*, among which thirteen pairs exhibited positive correlations while three pairs showed negative correlations. Overall, the co-expression analysis of the selected *EbWRKY* genes revealed that these genes might play an essential role in flavonoid synthesis.

3.8 Quantitative real-time PCR profiling characterization of genes under exogenous hormone treatment

To investigate the expression pattern responses of genes under ABA, SA, and GA3 exogenous hormones in *E. breviscapus*, eleven structural genes of the flavonoid biosynthesis pathway and five WRKY genes (*EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44*) were selected for qRT-PCR analysis after exogenous hormone treatment. Eleven genes involved in the flavonoid synthesis pathway exhibited significant up-regulation, implying their potential functional role in this biological process. (Figure 7; Supplementary Table S9). The relative expression of the selected key genes exhibited distinct temporal patterns in response to different treatments. Notably, *FLS* displayed the highest expression level, with a 20-fold increase observed after 4 hours of GA3 treatment and a 10-fold increase after 4 hours of ABA and SA treatment, followed by a subsequent decrease at 12 hours. *C4H*, *F6H*, and *F3H* showed similar expression patterns after 4 h of treatment, indicating that these genes are sensitive to GA3, SA, and

ABA. *CHI* and *PAL* showed relatively high expression levels after 4 h of ABA treatment (> 3.5-fold). *CHS* exhibited the highest expression level, with an 8-fold expression at 4 h of ABA treatment and a 6.8-fold expression with GA3 treatment. *FSH* showed a 2.9-fold higher expression after 4 h of SA treatment.

The expression patterns of *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* transcription factors related to the structural genes involved in flavonoid biosynthesis varied under different hormone treatments. *EbWRKY11* showed the highest expression level after 4 h of hormone treatment and gradually decreased after 12 h, indicating that *EbWRKY11* was sensitive to ABA, SA, and GA3. *EbWRKY30* was sensitive to ABA and had the highest expression level at 4 h with a 4.3-fold increase. In the SA treatment, an increase was followed by a decrease in the volatility of *EbWRKY30*. *EbWRKY31* showed a significant response to ABA treatment, and the expression level gradually decreased after SA and GA3 treatment at 4h. *EbWRKY36* exhibited a > 2-fold change in expression after 4 h of hormone treatments. *EbWRKY44* was more sensitive to SA than to ABA and GA3, with a 2.2-fold expression at 4 h. However, the gene expression level was highest at 12 h and 24 h of ABA treatment.

3.9 Protein-DNA interactions between EbWRKY11, EbWRKY36, EbWRKY44, and F7GAT and EbF6H

The yeast one-hybrid assay was conducted to validate the interaction between *EbWRKY36*, *EbWRKY44*, *EbWRKY11* and two key structural genes, *F7GAT* and *EbF6H*, which encode key enzymes involved in the conversion of apigenin to scutellarin. The

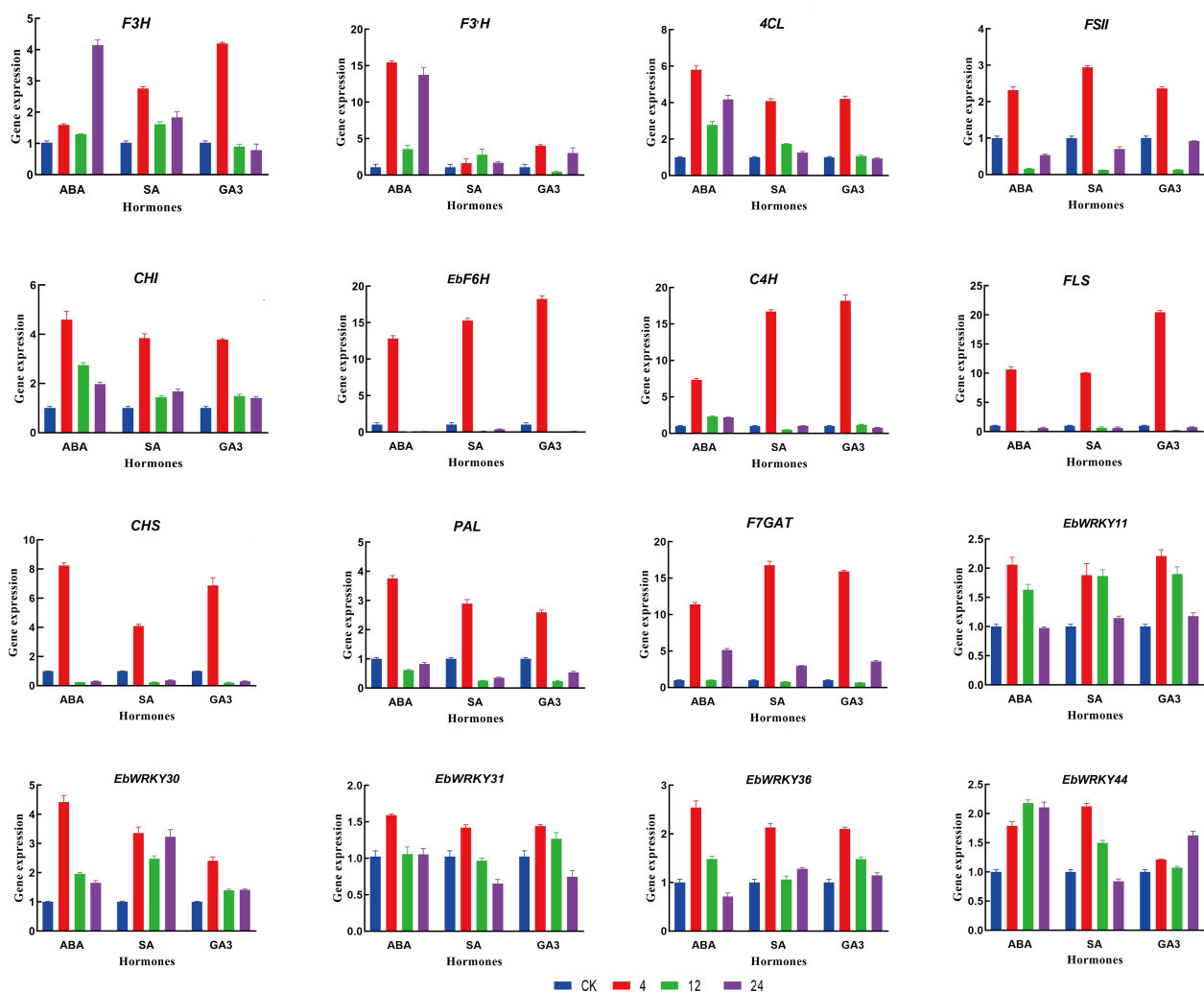


FIGURE 7

Relative expression of selected *Eb* genes in response to exogenous hormone treatment. Genes expression was analyzed by RT-qPCR. Blue was used as the untreated control (expression = 1); Red, green, and purple represent 4h, 12h, and 24h. Error bars represent standard errors. Data were calculated using the $2^{-\Delta\Delta C_t}$ method.

bait vector was constructed by utilizing the high GC% content of the *F7GAT* and *EbF6H* promoter domains. The results demonstrated that *EbWRKY36*, *EbWRKY44*, and *EbWRKY11* exhibit binding affinity towards the promoter region *F7GAT* (1801-2500bp) (Figure 8A). Unfortunately, *EbF6H* was proven to have a self-activation function (Figure 8B). The two predicted regions with high scores containing WRKY-binding sites (ATAGTCAACT and TTCAAAGTCAAA) were truncated for verification. Notably, self-activation was observed in the 1501-2100bp region, while the 501-800bp region exhibited no self-activation but lacked interaction with the transcription factors *EbWRKY36* and *EbWRKY44*, as well as *EbWRKY11* (Figures 8A, B; Supplementary Table S10). These findings confirm that these three WRKY transcription factors of *E. breviscapus* play a role in the transcriptional regulation of key structural genes and regulated biosynthesis of scutellarin, a crucial active ingredient.

4 Discussion

The origins of WRKY transcription factors can be traced back to prokaryotes, with their presence limited to certain diplomonads, social amoebae, other amoebozoan species, and members of the fungal class incertae sedis (Rinerson et al., 2015; Chen et al., 2017). The 75 *EbWRKYs* were classified into two major branches: Group I was separated into a single branch, while the remaining *EbWRKYs* formed two complex branches consisting of individual sub-branches including Group IIc, Group IIa + IIb, Group IId + IIe, and Group III. Group II has the most numbers of *EbWRKYs*. Based on previous studies, WRKYs might evolve from the common ancestors, Group I WRKYs representing a more primitive form that subsequently evolved into Group II, encompassing subgroups IIa + IIb, IIc, and IId + IIe, Group III was closely related to Group IId and Group IIe (Zhang and Wang, 2005). In contrast, the present

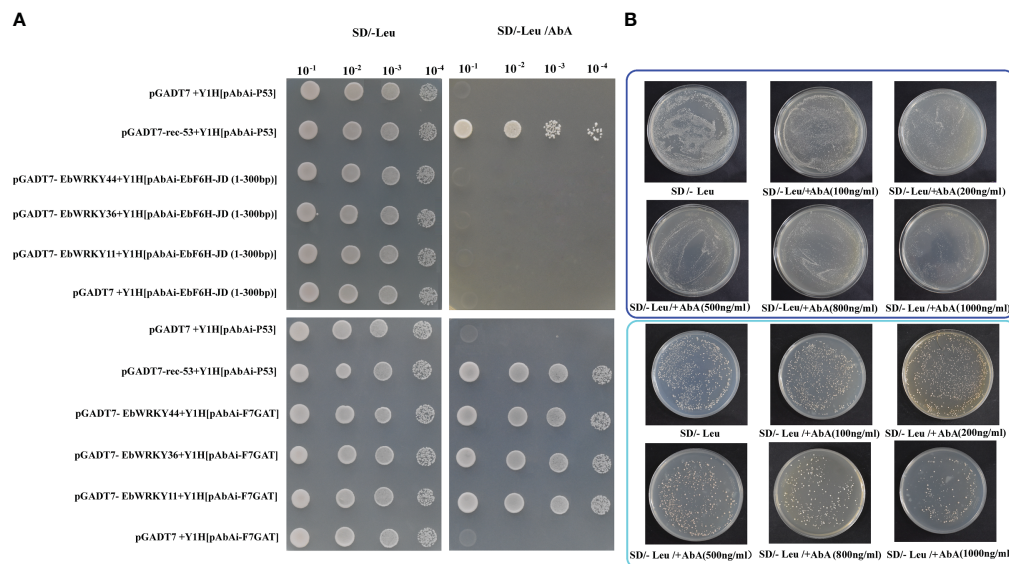


FIGURE 8

Protein-DNA interactions between EbWRKY11, EbWRKY36, EbWRKY44 with *EbF6H* and *F7GAT*. (A) Yeast One-Hybrid experiments. The pGADT7 was an AD empty vector; pGADT7-rec-53+Y1H[pAbAi-P53] was a positive control and pGADT7 +Y1H[pAbAi-P53] was a negative control. The *EbF6H*-JD (1-300bp) represents the promoter area of *EbF6H* 501-800bp. (B) *EbF6H* self-activation experiments. The upper blue box shows the full-length *EbF6H* promoter and the cyan box shows the area of 1501~2100bp promoter after the truncated *EbF6H*.

study revealed a closer genetic relationship between Group IIa + IIb and Group III compared to that between Group IIc + IIe, suggesting a stronger evolutionary connection of *EbWRKYs* in Group IIa + IIb with those in Group III. These findings imply a shared ancestral origin among these genes.

The conserved domains of the *EbWRKY* protein were further evaluated according to the motif characteristics of three categories and five subtypes. All WRKY protein sequences exhibited a completely conserved domain (WRKYGQK) and zinc finger structure associated with the WRKY motif except for *EbWRKY8*. The gene family phylogenetic analyses were consistent with the results of the motif structure and sequence alignment. The findings further substantiate that the evolutionary pattern of Group I WRKYs exhibits a higher degree of conservatism compared to other types, which was reported in previous study (Eulgem et al., 2000). Multiple sequence alignments indicated that, regardless of the common WRKYGQK heptapeptide sequence, there were other variations, mainly distributed in Group IIc (*EbWRKY68/69*) and Group IId (*EbWRKY2/72/73*). In addition, the WRKY domain variations of *EbWRKYs* occurred in Group III (*EbWRKY40*). The differences in the conserved domain of the WRKY protein may be caused by variations in the WRKYGQK heptapeptide sequence and zinc finger structure during evolution or the deletion mutation of amino acid residues (Zhang and Wang, 2005; Wang et al., 2011; Llorca et al., 2014). Mutations in the conserved domain reflect the diversity of the evolution of the plant WRKY gene family, similar minor variations have been observed in *Citrus* and *rice* (Xie et al., 2005; Ayadi et al., 2016). Variations in WRKYGQKs affect its affinity for the W-box and further influence its function (Eulgem et al., 2000; Maeo et al., 2001). Therefore, the interactions between *EbWRKY* proteins with these variations and downstream target

genes, and their binding preferences with cis-acting W-box elements, should be further investigated.

The bioactive flavonoids, particularly scutellarin, are predominantly distributed in the leaves of *E. breviscapus*, which serves as the primary raw material for pharmaceutical extraction. In this study, 54 flavone and flavonol compounds showed spatiotemporal accumulation in the leaves of *E. breviscapus* after hormone treatments, notably, GA3 significantly improved the accumulation of flavones and flavonols in the leaves. Meanwhile, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* exhibited specific high expression levels in the leaves of *E. breviscapus*, while the expression level of *EbWRKY11* was lower than that in the roots but higher than in other tissues after hormone treatment. *EbWRKY44* and *EbWRKY36* exhibit significant increases in response to three hormones and treatment for 4 hours, while *EbWRKY30* and *EbWRKY31* specifically respond to ABA treatment only, displaying an inverse expression pattern of transcripts. These tissue-specific expression patterns suggest the potential involvement of these five *EbWRKY* transcription factors in flavonoid metabolism within leaves, while also indicating a possible role for *EbWRKY11* in root metabolism and development.

The WRKY transcription factor can bind to the promoter regions of functional genes or be induced by external stimuli, thereby regulating gene transcription levels involved in secondary metabolite accumulation (Liu et al., 2015; Bray, 1997; Shinozaki and Yamaguchi, 2000). Our previous work showed that *PAL*, *EbF6H*, *C4H*, *4CL*, *F7GAT*, *CHI*, *CHS*, *FLS*, *FSII*, *F3H*, and *F3'H* were key structural genes regulating flavonoid biosynthesis in the leaves of *E. breviscapus* (Gao et al., 2022; Zhao et al., 2022). *MdWRKY11* regulates anthocyanin synthesis through directly binding to the flavonoid 3-O-glycosyl transferase promoter in apple (Liu et al.,

2019). *McWRKY71* controls *McANR* and proanthocyanidin synthesis in *Malus crabapple* (Zhang et al., 2022). Other reports demonstrated that *FaWRKY71* stimulates anthocyanin accumulation in strawberry (*Fragaria × ananassa*) by up-regulating *FaF3'H*, *FaLAR*, and *FaANR* (Yue et al., 2022). In this study, the structural genes of flavonoid biosynthesis of *E. breviscapus* significantly increased after 4 h of exogenous hormone treatment. The 11 structural gene expression patterns are basically consistent with the experimental results of qRT-PCR analysis in hormone treatment. *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* are closely related to the genes of flavonoid biosynthesis. qRT-PCR analysis further verified the expression patterns of the structural genes involved in the flavonoid biosynthesis pathway are consistent with *EbWRKYs*. Therefore, these five *EbWRKY* transcription factors may participate in the transcription of key structural genes that regulate flavonoid metabolite accumulation.

The phylogenetic analysis revealed that *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* were distributed across all three WRKY groups. The homologous genes of *EbWRKY44* in Group IId were *AtWRKY7/11/15/17* in *Arabidopsis*. In Group IIb, *AtWRKY6*, *AtWRKY31*, and *AtWRKY42* were co-orthologous to *EbWRKY31*. *EbWRKY30* in Group III was homologous to *AtWRKY30/46/41/53*. *EbWRKY36* and *EbWRKY11* were orthologs of *AtWRKY1* and *AtWRKY32*, which belonged to different nodes in the same branch of Group I. Previous studies have shown that WRKY transcription factors in *Arabidopsis* are widely involved in the regulation of biotic and abiotic stress, plant growth, and development. *AtWRKY15* regulates *Arabidopsis* growth and the salt stress response, whereas *AtWRKY7/11/17* are negative regulators of the PAMP immune system, which could enhance plant resistance to pathogens (Vanderauwera et al., 2012; Arraño-Salinas et al., 2018). In addition, *AtWRKY7* contains a CaM-binding domain, a new CaM-binding transcription factor that regulates plant growth and development and plays an important role in Ca²⁺ signal transduction (Park et al., 2005). Moreover, *MxWRKY55* and *VvWRKY28* from *Malus xiaojinensis* and grape respectively, belong to Group II with WRKY TFs playing a role in plant resistance and contributing to higher salt tolerance (Han et al., 2020; Liu et al., 2022). *EbWRKY31* and *EbWRKY44* both belong to Group II and may be involved in salt stress response. Overexpression of *AtWRKY30* enhances abiotic stress tolerance in *A. thaliana* at the early growth stage (Scarpeci et al., 2013). *AtWRKY1* and *AtWRKY41* can resist *Pseudomonas syringae* (Mukhi et al., 2021). *AtWRKY42* regulates Pi homeostasis to adapt to environmental changes, and *AtWRKY6* participates in Pi transportation (Robatzek and Somssich, 2001; Su et al., 2015). *AtWRKY53* regulates stomatal movement and negatively regulates drought resistance, whereas *AtWRKY46* is involved in the sensitivity of *Arabidopsis* to drought and salt stress (Ma et al., 2017; Freeborough et al., 2021). Based on the evolutionary analysis results of *EbWRKYs* and *AtWRKYs*, it was speculated that *EbWRKYs* homologous to various branches of *Arabidopsis* might participate in plant growth, development, and responses to biological and abiotic stresses with other members in the branch.

Transcription factors can act alone or in conjunction with other proteins to form complex binding complexes at gene promoters, thereby exerting control over gene expression through physical interactions that span long distances and coordinate the transcriptional activation of specific genes (Banerji et al., 1981; Karin, 1990; Latchman, 1997; Yao et al., 2015). In apple, *MdWRKY1* increases anthocyanin accumulation by activating *MdLNC499* and *MDERF109* expression (Ma et al., 2021). The interaction between *PyWRKY26* and *PybHLH3* targeting the promoter of *PyMYB114* may potentially modulate the accumulation of anthocyanins in red-skinned pears (Li et al., 2020). In this study, protein-DNA interactions between *EbWRKY11*, *EbWRKY36*, *EbWRKY44*, and *F7GAT* and *EbF6H*, which were typical glycosylase and hydroxylase involved in scutellarin biosynthesis in *E. breviscapus*. The results revealed that *EbWRKY36*, *EbWRKY44*, and *EbWRKY11* demonstrate direct binding to the promoter of *F7GAT*, whereas these three *EbWRKYs* did not exhibit interaction with the promoter region of *EbF6H* (501-800bp). *EbF6H* cannot be directly regulated by *EbWRKY36*, *EbWRKY44*, and *EbWRKY11*, however, it has been demonstrated in grape that the indirect regulation of structure genes on flavonoid biosynthesis hydroxylation occurs through the regulation of other regulatory elements. *VvWRKY26* is preferentially recruited by a *VvMYB5a*-driven MBW complex to regulate flavonoid hydroxylation (Amato et al., 2019).

5 Conclusion

In this study, a total of 75 *EbWRKY* transcription factors were predicted from the genome of *E. breviscapus*. The amino acid number, molecular weight, predicted isoelectric point (PI) value, chromosome position, domain pattern, and conservative motif of *EbWRKYs* were revealed by bioinformatics-based analyses. The specificity of *EbWRKYs* gene expression in different tissues and their expression pattern under hormone treatment were determined based on RNA sequencing. Combining metabolome and transcriptome results revealed the regulatory mechanism between the WRKY transcription factor and key genes involved in flavonoid biosynthesis. The expression patterns of *EbWRKY11*, *EbWRKY30*, *EbWRKY31*, *EbWRKY36*, and *EbWRKY44* transcription factors were similar to those of the 11 key structural genes involved in flavonoid biosynthesis. *EbWRKY36*, *EbWRKY44*, and *EbWRKY11* can interact with the promoter of *F7GAT*, which was the key glycosyltransferase involved in scutellarin biosynthesis. We provided comprehensive information about the WRKY gene family of *E. breviscapus* and the mechanism of *EbWRKY* genes involved in flavonoid metabolism regulation.

Data availability statement

The raw data of *E. breviscapus* transcriptome in the current study are available in the National Center for Biotechnology Information (NCBI) database under project number PRJNA971382 (<https://www.ncbi.nlm.nih.gov/bioproject/>

PRJNA971382). The genomic data of *E. breviscapus* is downloaded from the Medicinal Plants multi-Omics Database (<http://medicinalplants.ynau.edu.cn/genome/detail/68>).

Ethics statement

Experimental research and field studies on plants comply with relevant institutional, national, and international guidelines and legislation, and all methods were performed according to the relevant guidelines and regulations. The cultivated *E. breviscapus* were collected with official permissions of Yunnan Hongling Biological Technology Co., LTD. Honghe, China.

Author contributions

WS: Formal analysis, Writing – original draft. SZ: Data curation, Writing – original draft. QL: Software, Writing – review & editing. GX: Software, Writing – review & editing. YZ: Formal analysis, Project administration, Writing – review & editing. FW: Methodology, Writing – review & editing. GZ: Methodology, Writing – review & editing. SY: Funding acquisition, Investigation, Writing – review & editing. BH: Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by Fundamental Research Project of Yunnan (202101AS070037). The Science and Technique Programs in

Yunnan Province (202102AE090042). Yunnan Characteristic Plant Extraction Laboratory (2022YKZY001), the First Projects of Science and Technology Plan in the Biomedical field in 2021 (202102AA310048).

Acknowledgments

The authors thank the lab members for their assistance.

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.2024.1412574/full#supplementary-material>

References

- Agati, G., Azzarello, E., Pollastri, S., and Tattini, M. (2012). Flavonoids as antioxidants in plants: location and functional significance. *Plant science*. 196, 67–76. doi: 10.1016/j.plantsci.2012.07.014
- Amato, A., Cavallini, E., Walker, A. R., Pezzotti, M., Bliet, M., Quattrocchio, F., et al. (2019). The MYB5-driven MBW complex recruits a WRKY factor to enhance the expression of targets involved in vacuolar hyper-acidification and trafficking in grapevine. *Plant J*. 99, 1220–1241. doi: 10.1111/tj.14419
- Amato, A., Cavallini, E., Zenoni, S., Finezzo, L., Begheldo, M., Ruperti, B., et al. (2017). A Grapevine TTG2-Like WRKY transcription factor is involved in regulating vacuolar transport and flavonoid biosynthesis. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01979
- Arraño-Salinas, P., Domínguez-Figueroa, J., Herrera-Vásquez, A., Zavala, D., Medina, J., Vicente-Carbajosa, J., et al. (2018). WRKY7, -11 and -17 transcription factors are modulators of the bZIP28 branch of the unfolded protein response during PAMP-triggered immunity in *Arabidopsis thaliana*. *Plant Sci.* 277, 242–250. doi: 10.1016/j.plantsci.2018.09.019
- Ayadi, M., Hanana, M., Kharrat, N., Merchaoui, H., Marzoug, R. B., Lauvergeat, V., et al. (2016). The WRKY transcription factor family in *Citrus*: valuable and useful candidate genes for *Citrus* breeding. *Appl. Biochem. Biotechnol.* 180, 516–543. doi: 10.1007/s12010-016-2114-8
- Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*. 27, 299–308. doi: 10.1016/0092-8674(81)90413-X
- 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
- Bray, E. A. (1997). Plant responses to water deficit. *Trends Plant Sci.* 2, 48–54. doi: 10.1016/S1360-1385(97)82562-9
- Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., and He, Y. (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, F., Hu, Y., Vannozzi, A., Wu, K., Cai, H., Qin, Y., et al. (2017). The WRKY transcription factor family in model plants and crops. *Crit. Rev. Plant Sci.* 36, 311–335. doi: 10.1080/07352689.2018.1441103
- Chen, Y. J., Chen, C., Li, M. Y., Li, Q. Q., Zhang, X. J., Huang, R., et al. (2021). Scutellarin reduces cerebral ischemia reperfusion injury involving in vascular endothelium protection and PKG signal. *Nat. Prod. Bioprospect.* 11, 659–670. doi: 10.1007/s13659-021-00322-z
- Deng, Z. L., Münch, P. C., Mreches, R., and McHardy, A. C. (2022). Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.* 50, e60. doi: 10.1093/nar/gkac112
- Di, P., Wang, P., Yan, M., Han, P., Huang, X., Yin, L., et al. (2021). Genome-wide characterization and analysis of WRKY transcription factors in *Panax ginseng*. *BMC Genomics* 22, 1–15. doi: 10.1186/s12864-021-08145-5
- Dong, J., Chen, C., and Chen, Z. (2003). Expression profiles of the Arabidopsis WRKY gene superfamily during plant defense response. *Plant Mol. Biol.* 51, 21–37. doi: 10.1023/A:1020780022549
- Duan, S., Wang, J., Gao, C., Jin, C., Li, D., Peng, D., et al. (2018). Functional characterization of a heterologously expressed *Brassica napus* WRKY41-1 transcription factor in regulating anthocyanin biosynthesis in *Arabidopsis thaliana*. *Plant Sci.* 268, 47–53. doi: 10.1016/j.plantsci.2017.12.010

- Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., et al. (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* 6, 1060–1083. doi: 10.1038/nprot.2011.335
- Eulgem, T., Rushton, P. J., Robatzek, S., and Somssich, I. E. (2000). The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* 5, 199–206. doi: 10.1016/S1360-1385(00)01600-9
- Eulgem, T., Rushton, P. J., Schmelzer, E., Hahlbrock, K., and Somssich, I. E. (1999). Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *EMBO J.* 18, 4689–4699. doi: 10.1093/emboj/18.17.4689
- Freeborough, W., Gentle, N., and Rey, M. E. C. (2021). WRKY transcription factors in cassava contribute to regulation of tolerance and susceptibility to cassava mosaic disease through stress responses. *Viruses*. 13 (9), 1820. doi: 10.3390/v13091820
- Gao, Q., Song, W., Li, X., Xiang, C., Chen, G., Xiang, G., et al. (2022). Genome-wide identification of bHLH transcription factors: Discovery of a candidate regulator related to flavonoid biosynthesis in *Erigeron breviscapus*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.977649
- Grunewald, W., Smet, I., Lewis, D. R., Löffke, C., Jansen, L., Goeminne, G., et al. (2012). Transcription factor WRKY23 assists auxin distribution patterns during Arabidopsis root development through local control on flavonol biosynthesis. *Proc. Natl. Acad. Sci. U S A.* 109, 1554–1559. doi: 10.1073/pnas.1121134109
- Guo, C., Guo, R., Xu, X., Gao, M., Li, X., Song, J., et al. (2014). Evolution and expression analysis of the grape (*Vitis vinifera* L.) WRKY gene family. *J. Exp. Bot.* 65 (6), 1513–1528. doi: 10.1093/jxb/eru007
- Han, D., Ding, H., Chai, L., Liu, W., Zhang, Z., Hou, Y., et al. (2018a). Isolation and characterization of *MbWRKY1*, a WRKY transcription factor gene from *Malus baccata* (L.) Borkh involved in drought tolerance. *Can. J. Plant Sci.* 98, 1023–1034. doi: 10.1139/cjps-2017-0355
- Han, D., Hou, Y., Ding, H., Zhou, Z., Li, H., and Yang, G. (2018b). Isolation and preliminary functional analysis of *MbWRKY4* gene involved in salt tolerance in transgenic tobacco. *Int. J. Agric. Biol.* 20, 433–441. doi: 10.1080/17429145.2018.1499145
- Han, D., Zhou, Z., Du, M., Li, T., Wu, X., Yu, J., et al. (2020). Overexpression of a *Malus xiao*ensis* WRKY transcription factor gene (*MxWRKY55*) increased iron and high salinity stress tolerance in *Arabidopsis thaliana*. *In Vitro Cell. Dev. Biology-Plant* 56, 600–609. doi: 10.1007/s11627-020-10129-1
- Ishiguro, S., and Nakamura, K. (1994). Characterization of a cDNA encoding a novel DNA-binding protein, SPF1, that recognizes SP8 sequences in the 5' Upstream regions of genes coding for sporamin and β -amylase from sweet potato. *Mol. Gen. Genet.* 244, 563–571. doi: 10.1007/BF00282746
- Jin, J., Tian, F., Yang, D. C., Meng, Y. Q., Kong, L., Luo, J., et al. (2017). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi: 10.1093/nar/gkw982
- Johnson, C. S., Kolevski, B., and Smyth, D. R. (2002). TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of *Arabidopsis*, encodes a WRKY transcription factor. *Plant Cell*. 14, 1359–1375. doi: 10.1105/tpc.001404
- Ju, S. H., Tan, L. R., Liu, P. W., Tan, Y. L., Zhang, Y. T., Li, X. H., et al. (2021). Scutellarin regulates osteoarthritis *in vitro* by inhibiting the PI3K/AKT/mTOR signaling pathway. *Mol. Med. Rep.* 23, 83. doi: 10.3892/mmr.2020.11722
- Karin, M. (1990). Too many transcription factors: Positive and negative interactions. *New Biol.* 2, 126–131.
- Khan, M. I., Fatma, M., Per, T. S., Anjum, N. A., and Khan, N. A. (2015). Salicylic acid-induced abiotic stress tolerance and underlying mechanisms in plants. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00462
- Latchman, D. S. (1997). Transcription factors: An overview. *Int. J. Biochem. Cell Biol.* 29, 1305–1312. doi: 10.1016/S1357-2725(97)00085-X
- Li, Z., and Ahammed, G. J. (2023). Hormonal regulation of anthocyanin biosynthesis for improved stress tolerance in plants. *Plant Physiol. Biochem.* 201, 107835. doi: 10.1016/j.plaphy.2023.107835
- Li, C., Li, D., Shao, F., and Lu, S. (2015). Molecular cloning and expression analysis of WRKY transcription factor genes in *Salvia miltiorrhiza*. *BMC Genomics* 16, 200. doi: 10.1186/s12864-015-1411-x
- Li, C., Wu, J., Hu, K. D., Wei, S. W., Sun, H. Y., Hu, L. Y., et al. (2020). *PyWRKY26* and *PyHLLH3* cotargeted the *PyMYB114* promoter to regulate anthocyanin biosynthesis and transport in red-skinned pears. *Hortic. Res.* 7, 37. doi: 10.1038/s41438-020-0254-z
- Liu, X., Cheng, J., Zhang, G., Ding, W., Duan, L., and Yang, J. (2018). Engineering yeast for the production of breviscapine by genomic analysis and synthetic biology approaches. *Nat. Commun.* 9, 448. doi: 10.1038/s41467-018-02883-z
- Liu, W., Liang, X., Cai, W., Wang, H., Liu, X., Cheng, L., et al. (2022). Isolation and functional analysis of *VvWRKY28*, a *Vitis vinifera* WRKY transcription factor gene, with functions in tolerance to cold and salt stress in transgenic *Arabidopsis thaliana*. *Int. J. Mol. Sci.* 23, 13418. doi: 10.3390/ijms232113418
- Liu, J., Osbourn, A., and Ma, P. (2015). MYB transcription factors as regulators of phenylpropanoid metabolism in plants. *Mol. Plant* 8, 689–708. doi: 10.1016/j.molp.2015.03.012
- Liu, W. J., Wang, Y. C., Yu, L., Jiang, H. Y., Guo, Z. W., Xu, H. F., et al. (2019). *MdWRKY11* participates in anthocyanin accumulation in red-fleshed apples by affecting MYB transcription factors and the photoresponse factor *MdHY5*. *J. Agric. Food Chem.* 67, 8783–8793. doi: 10.1021/acs.jafc.9b02920
- Llorca, C. M., Potschin, M., and Zentgraf, U. (2014). bZIPs and WRKYs: two large transcription factor families executing two different functional strategies. *Front. Plant Sci.* 5. doi: 10.3389/fpls.2014.00169
- Lucho-Constantino, G. G., Zaragoza-Martínez, F., Ponce-Noyola, T., Carlos, M., Cerda-García, R., and Ramos-Valdivia, A. C. (2017). Antioxidant responses under jasmonic acid elicitation comprise enhanced production of flavonoids and anthocyanins in *Jatropha curcas* leaves. *Acta Physiol. Plant* 39, 165. doi: 10.1007/s11738-017-2461-2
- Ma, J., Gao, X., Liu, Q., Shao, Y., Zhang, D., Jiang, L., et al. (2017). Overexpression of *TaWRKY146* increases drought tolerance through inducing stomatal closure in *Arabidopsis thaliana*. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.02036
- Ma, H., Yang, T., Li, Y., Zhang, J., Wu, T., Song, T., et al. (2021). The long noncoding RNA *MdLNC499* bridges *MdWRKY1* and *MdERF109* function to regulate early-stage light-induced anthocyanin accumulation in apple fruit. *Plant Cell*. 33, 3309–3330. doi: 10.1093/plcell/koab188
- Maeo, K., Hayashi, S., Kojima-Suzuki, H., Morikami, A., and Nakamura, K. (2001). Role of conserved residues of the WRKY domain in the DNA-binding of tobacco WRKY family proteins. *Biosci. Biotechnol. Biochem.* 65, 2428–2436. doi: 10.1271/bbb.65.2428
- Mukhi, N., Brown, H., Gorenkin, D., Ding, P., Bentham, A. R., Stevenson, C. E. M., et al. (2021). Perception of structurally distinct effectors by the integrated WRKY domain of a plant immune receptor. *Proc. Natl. Acad. Sci. U S A.* 118 (50), e2113996118. doi: 10.1073/pnas.2113996118
- Park, C. Y., Lee, J. H., Yoo, J. H., Moon, B. C., Choi, M. S., Kang, Y. H., et al. (2005). WRKY group IId transcription factors interact with calmodulin. *FEBS Lett.* 579, 1545–1550. doi: 10.1016/j.febslet.2005.01.057
- Pourcel, L., Routaboul, J. M., Cheymier, V., Lepiniec, L., and Debeaujon, I. (2007). Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* 12, 29–36. doi: 10.1016/j.tplants.2006.11.006
- Rinerson, C. I., Rabara, R. C., Tripathi, P., Shen, Q. J., and Rushton, P. J. (2015). The evolution of WRKY transcription factors. *BMC Plant Biol.* 15, 66. doi: 10.1186/s12870-015-0456-y
- Robatzek, S., and Somssich, I. E. (2001). A new member of the *Arabidopsis* WRKY transcription factor family, *AtWRKY6*, is associated with both senescence- and defence-related processes. *Plant J.* 28, 123–133. doi: 10.1046/j.1365-3113.2001.01131.x
- Rushton, P. J., Somssich, I. E., Ringler, P., and Shen, Q. J. (2010). WRKY transcription factors. *Trends Plant Sci.* 15, 247–258. doi: 10.1016/j.tplants.2010.02.006
- Scarpeci, T. E., Zanon, M. I., Mueller-Roeber, B., and Valle, E. M. (2013). Overexpression of *AtWRKY30* enhances abiotic stress tolerance during early growth stages in *Arabidopsis thaliana*. *Plant Mol. Biol.* 83, 265–277. doi: 10.1007/s11103-013-0090-8
- Schluttenhofer, C., and Yuan, L. (2015). Regulation of specialized metabolism by WRKY transcription factors. *Plant Physiol.* 167, 295–306. doi: 10.1104/pp.114.251769
- Shinozaki, K., and Yamaguchi, K. (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.* 3, 217–223. doi: 10.1016/S1369-5266(00)80068-0
- Su, T., Xu, Q., Zhang, F. C., Chen, Y., Li, L. Q., Wu, W. H., et al. (2015). *WRKY42* modulates phosphate homeostasis through regulating phosphate translocation and acquisition in *Arabidopsis*. *Plant Physiol.* 167, 1579–1591. doi: 10.1104/pp.114.253799
- Vanderauwera, S., Vandenbroucke, K., Inzé, A., Cotte, B., Mühlenbock, P., Rycke, R., et al. (2012). *AtWRKY15* perturbation abolishes the mitochondrial stress response that steers osmotic stress tolerance in *Arabidopsis*. *Proc. Natl. Acad. Sci. U S A.* 109 (49), 20113–20118. doi: 10.1073/pnas.1217516109
- Vives-Peris, V., Marnaneu, D., Gómez-Cadenas, A., and Pérez-Clemente, R. M. (2018). Characterization of *Citrus* WRKY transcription factors and their responses to phytohormones and abiotic stresses. *Springer Netherlands*. 62, 33–44. doi: 10.1007/s10535-017-0737-4
- Wang, Y., Liu, X., Chen, B., Liu, W., Guo, Z., Liu, X., et al. (2022). Metabolic engineering of *Yarrowia lipolytica* for scutellarin production. *Synth Syst. Biotechnol.* 7, 958–964. doi: 10.1016/j.synbio.2022.05.009
- Wang, Z., Luo, Z., Liu, Y., Li, Z., Liu, P., Bai, G., et al. (2021). Molecular cloning and functional characterization of *NtWRKY11b* in promoting the biosynthesis of flavonols in *Nicotiana tabacum*. *Plant Sci.* 304, 110799. doi: 10.1016/j.plantsci.2020.110799
- Wang, J. B., Pu, S. B., Sun, Y., Li, Z. F., Niu, M., Yan, X. Z., et al. (2014). Metabolomic profiling of autoimmune hepatitis: the diagnostic utility of nuclear magnetic resonance spectroscopy. *J. Proteome Res.* 13, 3792–3801. doi: 10.1021/pr500462f
- Wang, N., Song, G., Zhang, F., Shu, X., Cheng, G., Zhuang, W., et al. (2023). Characterization of the WRKY gene family related to anthocyanin biosynthesis and the regulation mechanism under drought stress and methyl jasmonate treatment in *Lycoris radiata*. *Int. J. Mol. Sci.* 24, 2423. doi: 10.3390/ijms24032423

- Wang, Q., Wang, M., Zhang, X., Hao, B., Kaushik, S. K., and Pan, Y. (2011). WRKY gene family evolution in *Arabidopsis thaliana*. *Genetica*. 139, 973–983. doi: 10.1007/s10709-011-9599-4
- Wei, Y., Meng, N., Wang, Y., Cheng, J., Duan, C., and Pan, Q. (2023). Transcription factor VvWRKY70 inhibits both norisoprenoid and flavonol biosynthesis in grape. *Plant Physiol.* 193, 2055–2070. doi: 10.1093/plphys/kiad423
- Xie, Z., Zhang, Z. L., Zou, X., Huang, J., Ruas, P., Thompson, D., et al. (2005). Annotations and functional analyses of the rice WRKY gene superfamily reveal positive and negative regulators of abscisic acid signaling in aleurone cells. *Plant Physiol.* 137, 176–189. doi: 10.1104/pp.104.054312
- Xu, N., Liu, S., Lu, Z., Pang, S., Wang, L., Wang, L., et al. (2020). Gene expression profiles and flavonoid accumulation during salt stress in *Ginkgo biloba* seedlings. *Plants*. 9 (9), 1162. doi: 10.3390/plants9091162
- Yamamoto, R., Ma, G., Zhang, L., Hirai, M., Yahata, M., Yamawaki, K., et al. (2020). Effects of salicylic acid and methyl jasmonate treatments on flavonoid and carotenoid accumulation in the juice sacs of satsuma mandarin *in vitro*. *Appl. Sci.* 10 (24), 8916. doi: 10.3390/app10248916
- Yang, L., Luo, S., Jiao, J., Yan, W., Zeng, B., He, H., et al. (2023). Integrated transcriptomic and metabolomic analysis reveals the mechanism of gibberellic acid regulates the growth and flavonoid synthesis in *phellodendron chinense schneid* seedlings. *Int. J. Mol. Sci.* 24, 16045. doi: 10.3390/ijms242216045
- Yang, L., Tao, Y., Luo, L., Zhang, Y., Wang, X., and Meng, X. (2022a). Dengzhan Xixin injection derived from a traditional Chinese herb *Erigeron breviscapus* ameliorates cerebral ischemia/reperfusion injury in rats via modulation of mitophagy and mitochondrial apoptosis. *J. Ethnopharmacol.* 288, 114988. doi: 10.1016/j.jep.2022.114988
- Yang, L., Yan, Y., Zhao, B., Xu, H., Su, X., and Dong, C. (2022b). Study on the regulation of exogenous hormones on the absorption of elements and the accumulation of secondary metabolites in the medicinal plant *artemisia argyi* leaves. *Metabolites*. 12, 984. doi: 10.3390/metabo12100984
- Yang, Y., Zhou, Y., Chi, Y., Fan, B., and Chen, Z. (2017). Characterization of soybean wrky gene family and identification of soybean WRKY genes that promote resistance to soybean cyst nematode. *Sci. Rep.* 7, 17804. doi: 10.1038/s41598-017-18235-8
- Yao, L., Shen, H., Laird, P. W., Farnham, P. J., and Berman, B. P. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* 16, 105. doi: 10.1186/s13059-015-0668-3
- Yin, W., Wang, X., Liu, H., Wang, Y., Nocker, S., Tu, M., et al. (2022). Overexpression of VqWRKY31 enhances powdery mildew resistance in grapevine by promoting salicylic acid signaling and specific metabolite synthesis. *Hortic. Res.* 9, uhab064. doi: 10.1093/hr/uhab064
- Yu, Y., Hu, R., Wang, H., Cao, Y., He, G., Fu, C., et al. (2013). MlWRKY12, a novel Miscanthus transcription factor, participates in pith secondary cell wall formation and promotes flowering. *Plantsci* 212, 1–9. doi: 10.1016/j.plantsci.2013.07.010
- Yu, F., Huaxia, Y., Lu, W., Wu, C., Cao, X., Guo, X., et al. (2012). GhWRKY15, a member of the WRKY transcription factor family identified from cotton (*Gossypium hirsutum* L.), is involved in disease resistance and plant development. *BMC Plant Biol.* 12, 144. doi: 10.1186/1471-2229-12-144
- Yue, M., Jiang, L., Zhang, N., Zhang, L., Liu, Y., Wang, Y., et al. (2022). Importance of FaWRKY71 in strawberry (*Fragaria x ananassa*) fruit ripening. *Int. J. Mol. Sci.* 23, 12483. doi: 10.3390/ijms232012483
- Zeng, M., Zhong, Y., Guo, Z., Yang, H., Zhu, H., Zheng, L., et al. (2022). Expression and functional study of bcWRKY1 in *baphicacanthus cusia* (Nees) bremerk. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.919071
- Zhang, Y., and Wang, L. (2005). The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.* 5, 1. doi: 10.1186/1471-2148-5-1
- Zhang, J., Wang, Y., Mao, Z., Liu, W., Ding, L., Zhang, X., et al. (2022). Transcription factor McWRKY71 induced by ozone stress regulates anthocyanin and proanthocyanidin biosynthesis in *Malus crabapple*. *Ecotoxicol Environ. Saf.* 232, 113274. doi: 10.1016/j.ecoenv.2022.113274
- Zhao, Y., Zhang, G., Tang, Q., Song, W., Gao, Q., Xiang, G., et al. (2022). EbMYBPI, a R2R3-MYB transcription factor, promotes flavonoid biosynthesis in *Erigeron breviscapus*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.946827
- Zhu, Z. P., Yu, J. X., Liu, F. F., Zhu, D. W., Xiong, A. S., and Sun, M. (2023). AeWRKY32 from okra regulates anthocyanin accumulation and cold tolerance in *Arabidopsis*. *J. Plant Physiol.* 287, 154062. doi: 10.1016/j.jplph.2023.154062



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Dongyan Zhao,
Cornell University, United States
Diaga Diouf,
Cheikh Anta Diop University, Senegal
Xuming Li,
Hugo Biotechnologies Co., Ltd., China

*CORRESPONDENCE

Yang Dong
✉ loyangyang@163.com
Wen-Bin Yu
✉ yuwenbin@xtbg.ac.cn

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

RECEIVED 05 June 2024

ACCEPTED 16 July 2024

PUBLISHED 02 August 2024

CITATION

Chen B-Z, Li D-W, Luo K-Y, Jiu S-T, Dong X,
Wang W-B, Li X-Z, Hao T-T, Lei Y-H,
Guo D-Z, Liu X-T, Duan S-C, Zhu Y-F,
Chen W, Dong Y and Yu W-B (2024)
Chromosome-level assembly of *Lindenbergia
philippensis* and comparative genomic
analyses shed light on genome
evolution in Lamiales.
Front. Plant Sci. 15:1444234.
doi: 10.3389/fpls.2024.1444234

COPYRIGHT

© 2024 Chen, Li, Luo, Jiu, Dong, Wang, Li,
Hao, Lei, Guo, Liu, Duan, Zhu, Chen, Dong and
Yu. 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.

Chromosome-level assembly of *Lindenbergia philippensis* and comparative genomic analyses shed light on genome evolution in Lamiales

Bao-Zheng Chen^{1,2†}, Da-Wei Li^{2†}, Kai-Yong Luo^{1,2},
Song-Tao Jiu³, Xiao Dong², Wei-Bin Wang², Xu-Zhen Li²,
Ting-Ting Hao², Ya-Hui Lei², Da-Zhong Guo^{1,2}, Xu-Tao Liu^{1,2},
Sheng-Chang Duan², Yi-Fan Zhu^{1,2}, Wei Chen², Yang Dong^{2*}
and Wen-Bin Yu^{4,5*}

¹College of Food Science and Technology, Yunnan Agricultural University, Kunming, Yunnan, China,

²Yunnan Provincial Key Laboratory of Biological Big Data, Yunnan Agricultural University, Kunming, Yunnan, China, ³Department of Plant Science, School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai, China, ⁴Center for Integrative Conservation and Yunnan Key Laboratory for the Conservation of Tropical Rainforests and Asian Elephants, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan, China, ⁵Southeast Asia Biodiversity Research Institute, Chinese Academy of Sciences, Mengla, Yunnan, China

Lamiales, comprising over 23,755 species across 24 families, stands as a highly diverse and prolific plant group, playing a significant role in the cultivation of horticultural, ornamental, and medicinal plant varieties. Whole-genome duplication (WGD) and its subsequent post-polyploid diploidization (PPD) process represent the most drastic type of karyotype evolution, injecting significant potential for promoting the diversity of this lineage. However, polyploidization histories, as well as genome and subgenome fractionation following WGD events in Lamiales species, are still not well investigated. In this study, we constructed a chromosome-level genome assembly of *Lindenbergia philippensis* (Orobanchaceae) and conducted comparative genomic analyses with 14 other Lamiales species. *L. philippensis* is positioned closest to the parasitic lineage within Orobanchaceae and has a conserved karyotype. Through a combination of Ks analysis and syntenic depth analysis, we reconstructed and validated polyploidization histories of Lamiales species. Our results indicated that *Primulina huaijiensis* underwent three rounds of diploidization events following the γ -WGT event, rather than two rounds as reported. Besides, we reconfirmed that most Lamiales species shared a common diploidization event (L-WGD). Subsequently, we constructed the Lamiales Ancestral Karyotype (LAK), comprising 11 proto-chromosomes, and elucidated its evolutionary trajectory, highlighting the highly flexible reshuffling of the Lamiales paleogenome. We identified biased fractionation of subgenomes following the L-WGD event across eight species, and highlighted the positive

impacts of non-WGD genes on gene family expansion. This study provides novel genomic resources and insights into polyploidy and karyotype remodeling of Lamiales species, essential for advancing our understanding of species diversification and genome evolution.

KEYWORDS

Lindenbergia philippensis, polyploidization history, karyotype evolutionary trajectories, Lamiales, genome assembly

Introduction

Whole-genome duplication (WGD) or polyploidization is a prevalent process in terrestrial plants, contributing to genetic diversity, particularly in ferns and angiosperms (Julca et al., 2018; Kubis et al., 1998; Jurka, 2000; Korf, 2004; Majoros et al., 2004; Li and Durbin, 2009; Katoh and Standley, 2013; Kellogg, 2016; Li et al., 2016; Kalyaanamoorthy et al., 2017; Landis et al., 2018; Luo et al., 2018; Mandáková and Lysak, 2018; Li et al., 2019; Lovell et al., 2021; Kong et al., 2023; Liao et al., 2023; Liu et al., 2023; Letunic and Bork, 2024). An increasing number of WGD events have been identified across various lineages from whole-genomic sequencing and comparative genomic analyses (Cui et al., 2006; Soltis et al., 2009; Jiao et al., 2011; Vanneste et al., 2014; Van de Peer et al., 2017). WGD events generally arise through two primary mechanisms: autopolyploidization, involving whole-genome duplication within a single species, and allopolyploidization, resulting from the hybridization of two distinct species (Stebbins, 1947; Cheng et al., 2018). WGD events can provide their ancestors with a 'genomic playground', enabling new mutations to arise and tend to be fixed (through gene sub-functionalization and/or neofunctionalization). Consequently, these may contribute to physiological and morphological innovations, making WGD events as a significant driving force for species diversification and environmental adaptation (Cheng et al., 2018; Ren et al., 2018).

WGD events play important roles in promoting angiosperm diversification. However, whether these events are correlated with higher diversification rates remains a subject of debate (Tank et al., 2015; Kellogg, 2016; Landis et al., 2018). The 'lag phase' model, positing a delay between polyploidization events and subsequent lineage diversification, offers critical insights into influences of WGD events on species diversification (Dodsworth et al., 2015; Tank et al., 2015; Clark and Donoghue, 2017; Mandáková and Lysak, 2018). In other words, the WGD event likely initiated many speciation events across angiosperm lineages and also provided the genetic basis for the post-polyploid diploidization (PPD) process (Mandáková and Lysak, 2018). PPD process is different from WGD events by involving a process of karyotype evolutionary trajectories, which primarily includes changes in genome size, chromosomal rearrangements (alterations in chromosomal number and structure), subgenome-specific fractionation (including biased

gene retention/loss and gene sub-/neofunctionalization), differential expression of homologous genes, activation of transposable elements (TE), and epigenetic reprogramming (Paterson et al., 2004; Wang et al., 2005; Mandáková and Lysak, 2018; Zhuang et al., 2019). Therefore, the PPD process may also play a significant role in promoting the diversification rate of angiosperms.

The evolutionary mechanism and significance of promoting species diversity through the PPD process have been elucidated and reviewed by several studies (Mandáková et al., 2017; Mandáková and Lysak, 2018; Mayrose and Lysak, 2021). Generally, dysploid or non-dysploid changes in chromosome number and the fractionation of duplicated genes represent the primary aspects of the PPD process. Among them, chromosomal changes arising from dysploid alterations can radically increase or decrease the base number of chromosomes. Both descending and ascending dysploidy are significant in karyotype evolution, with the latter primarily observed in a few plant groups possessing monocentric chromosomes, such as the cycad genus *Zamia* (Rastogi and Ohri, 2019; Mayrose and Lysak, 2021). The evolution of land plant chromosomes is predominantly characterized by descending dysploidy (Carta et al., 2020; Mayrose and Lysak, 2021; Wang et al., 2022c; Kong et al., 2023). Centric fission is traditionally considered the most common form of ascending dysploidy (Birchler and Han, 2018). Unlike ascending dysploidy, descending dysploidy can be initiated by two mechanisms, including end-to-end joining (EEJ) and nested chromosome fusion (NCF) (Morin et al., 2017; Ren et al., 2019; Sun et al., 2022). In general, chromosomal diploidization can also be accompanied by various non-dysploid chromosomal rearrangements (CRs), such as inversions, reciprocal translocations, deletions, and duplications (Schubert and Lysak, 2011; Sun et al., 2022). With the alterations in dysploidy and non-dysploidy, the karyotype of specific lineages will undergo significant reshuffling, leading to karyotype modifications and potentially initiating interspecific reproductive barriers. Consequently, these processes may enable some species to acquire evolutionarily advantageous genetic diversity, thus adapting to a changing environment (Soltis et al., 2009; Clark and Donoghue, 2017).

In addition to dysploid or non-dysploid changes, the prevalence of dominant subgenomes, resulting from the preferential retention

of genes, is notable in many lineages that have undergone WGD events (Edger et al., 2017; Lovell et al., 2021; Wang et al., 2022b). Consequently, compared to a submissive subgenome, a dominant subgenome often retains more ancestral genes, exhibits higher levels of homologous gene expression, and undergoes stronger purifying selection (Sun et al., 2023). The biased retention (fractionation) of redundant genes resulting from WGD events may facilitate the adaptation of lineage-specific species to diverse ecological environments during speciation. Wu et al. (2020), for example, investigated gene duplicates across 25 genomes, revealing that duplicates retained following WGD events often correlate with environmental adaptability. Specifically, gene families associated with cold and dark conditions were frequently preserved in several lineages following WGD events around the Cretaceous-Paleogene boundary, a period marked by significant global cooling and darkness. Benefiting from karyotype changes, lineage-specific species evolve towards advantageous genetic diversity through the PPD process. This evolutionary advantage provides them with greater buffering capacity against mutations than their ancestors, thereby aiding speciation and enhancing adaptability in harsh environments (Comai, 2005; Ren et al., 2018; Clo, 2022). However, elucidating the complex process of PPD is challenging because, in most species, the ancestral chromosome tend to scatter and fragment within the new karyotype due to changes following the long evolutionary history (Damas et al., 2018; Ren et al., 2019; Zhao et al., 2021). Consequently, the intricate process of PPD, which involves a range of evolutionary modifications, remains a largely overlooked and understudied topic, particularly in certain specific lineages.

Representing one of the most abundant and diverse plant groups, the order Lamiales comprises over 23,755 species and 24 families (<https://www.britannica.com/plant/Lamiales>). These plants play a crucial role in providing a wide variety of horticultural, ornamental, and medicinal species. Besides, a variety of ecotype plants can be found in this lineage, including autotrophic and heterotrophic (parasitic and carnivorous) plants, aquatic and terrestrial plants. The high species diversity in Lamiales can be directly reflected in the abundant genetic materials. More importantly, almost all Lamiales species shared a common WGD event (the *L event*), and most retain a relatively complete ancestral karyotype, making them as ideal resources for investigating the PPD process (Feng et al., 2020).

The history of polyploidization and the PPD process have long been subjects of extensive study due to their significant roles in species adaptation and evolution. However, research across many lineages has been limited by a lack of comprehensive genomic resources. Encouragingly, the increasing availability of chromosomal-level genome assemblies is now enabling more detailed investigations into the history of polyploidization and the evolutionary trajectories of karyotypes within specific lineages. Significant advances have been made in some specific lineages such as Asteraceae (Kong et al., 2023), Cucurbitaceae (Wang et al., 2022a), and Nyssaceae (Feng et al., 2024). *L. philippensis* is part of Orobanchaceae in Lamiales with a unique taxonomic status, being closest to the parasitic lineage within Orobanchaceae (Li et al.,

2019; Mutuku et al., 2021). Besides, *L. philippensis* exhibited a conserved karyotype according to our previously exploration. In this study, to provide more insightful information about the polyploidization history, karyotype evolutionary trajectories, and the subgenomes evolutionary traits in the Lamiales, we assembled a chromosome-level genome of *L. philippensis* using Oxford Nanopore Technology (ONT) sequencing, Illumina sequencing, and high-throughput chromosome conformation capture (Hi-C) technology. Furthermore, we conducted a comparative genomic analysis on *L. philippensis* and other 14 genomes from 12 families within the order Lamiales, with *Vitis vinifera* and *Ophiorrhiza pumila* as outgroup references. The polyploidization histories of most Lamiales genomes were validated and corrected through combined Ks and syntenic depth analyses. Additionally, an ancestral karyotype of Lamiales species was constructed, and its evolutionary trajectories were deciphered in eight Lamiales species. Our study provides valuable genomic resources and will facilitate further research into genome evolution and the PPD process in Lamiales.

Materials and methods

Plant materials and DNA extraction

The plant samples of *L. philippensis* were collected from the same adult plant cultivated at Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, and identified by Professor Wen-Bin Yu. Fresh leaves were stored in liquid nitrogen and sent to Novogene Co., Ltd. for sequencing (Beijing, China). The high-quality genomic DNA of *L. philippensis* was prepared by a modified CTAB method (Karoonthaisiri et al., 2020) and purified with QIAGEN® Genomic kit (QIAGEN, USA) at Novogene Co., Ltd. (Beijing, China). The quality and quantity of the extracted genomic DNA were assessed using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA), Qubit dsDNA HS Assay Kit on a Qubit 3.0 fluorometer (Life Technologies, Carlsbad, CA, USA) and electrophoresis on a 0.8% agarose gel, respectively.

Long read sequencing

For long-read sequencing, a total of 2 µg DNA was used for the ONT library construction. After the sample was qualified, long DNA fragments are selected using the BluePippin system (Sage Science, Beverly, MA, USA). Further, the ends of DNA fragments were repaired and a ligation reaction was conducted using the NEBNext® Ultra™ II End Repair/dA-Tailing Module Kit. The ONT library with an insert size of 30 kb was prepared using the ligation sequencing kit 1D (SQKLSK109; Oxford Nanopore Technologies, Oxford, UK) according to the manufacturer's instructions. The ONT sequencing was then performed on an Oxford Nanopore PromethION 48 platform at Novogene Co., Ltd. (Beijing, China).

Illumina short read sequencing

In total, 1 µg DNA was used as the input material and sequencing library was generated using the VAHTS Universal DNA Library Prep Kit for MGI (Vazyme, Nanjing, China). Following the manufacturer's recommendations, and index codes were added to attribute sequences to sample. The Library quantification and size were measured using Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA) and Bioanalyzer 2100 system (Agilent Technologies, CA, USA). A paired-end library was created with a 350 bp insert size using the GenElute Plant Genomic DNA Miniprep kits following the manufacturer's instructions (Sigma-Aldrich, Corp., St. Louis, MO, USA). Subsequently, the short-read library was performed on the Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA).

Hi-C library construction and sequencing

The Hi-C libraries were constructed following established protocols (Padmarasu et al., 2019). Initially, samples were cross-linked under vacuum infiltration using formaldehyde. Subsequently, the cross-linked samples were subsequently digested using *DpnII*. After reversing cross-links, the ligated DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) according to the manufacture's instruction. Purified DNA was then sheared to 300 bp to 500 bp fragments, which underwent blunt-end repair, A-tailing, and adaptor addition. The resulting fragments were purified through biotin-streptavidin-mediated pull-down and subjected to PCR amplification. Finally, the Hi-C libraries were quantified and sequenced on the Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA).

Genome assembly and quality evaluation

Prior to conducting the assembly, it is imperative to conduct a comprehensive survey of the genomic features. To accomplish this, we utilized clean paired-end short reads and employed GenomeScope (v2.0) and Jellyfish (v2.2.10) with default parameters to assess the genome size, heterozygosity, and repeat content of the *L. philippensis* genome (Marçais and Kingsford, 2011; Ranallo-Benavidez et al., 2020). Furthermore, flow cytometry (BD FACSCalibur) was also used to investigate the genome size. For the genome assembly, we initially assembled the clean long reads to generate the draft assembly using NextDenovo (v2.4.0) with the following parameters "task = all; rerun = 3; read_cuoff = 1k; seed_cutoff = 8k; seed_cuoff = 8k; genome_size = 400 m; seed_cutfiles = 80; blocksize = 10g; pa_correction = 80; minimap2_options_raw = -x ava-ont -t 16; sort_options = -m 10g -t 16 -k 50; correction_options = -p 32 random_round = 100 minimap2_options_cns = -x ava-ont -t 20 -k17 -w17; nextgraph_options = -a 1". Subsequently, the draft assembly underwent three rounds of polishing using NextPolish (v1.3.1) with the following parameters "rerun = 3; parallel_jobs = 8;

multithread_jobs = 8; sgs_options = -max_depth 100 -bwa". To obtain a preliminary genome assembly, haplotyped duplication sequences were filtered using Redundans (v1.01) with parameters "ident=0.95, ovl=0.95" (Pryszcz and Gabaldón, 2016). For scaffolding contigs, Hi-C data were mapped to the *L. philippensis* preliminary assembly using Juicer (v1.6.2) with parameters of "-s *DpnII* -t 40" (Durand et al., 2016). Subsequently, the valid reads were utilized to order and orient the contigs by employing 3D-DNA (Dudchenko et al., 2017). Any missing joins were rectified based on the Hi-C contact signals using Juicebox (v1.11.08) (<https://github.com/aidenlab/juicebox>). The completeness of the genome assembly was evaluated using BUSCO (v5.1.2) with "eukaryota_odb10" dataset downloaded from the BUSCO website (<https://busco-archive.ezlab.org/v3/>) (Seppey et al., 2019). We utilized BWA-MEM (v0.7.12) (Li and Durbin, 2009) for mapping Illumina reads to the assembly and computed mapping statistics with SAMtools (v1.9) using the "flagstat" module (Danecek et al., 2021).

For transcriptome assembly, we downloaded the raw reads of RNA sequencing data from NCBI (ERR2040586, ERR2040587) and used Fastp (v0.20.1) to filter the low quality reads with the following parameters "-q 30 -u 40 -l 50 -w 16". Trinity (v2.11.0) with the following parameters of "-seqType fq -JM 300G -CPU 20" was used to perform the transcriptome *de novo* assembly (Grabherr et al., 2011).

Genome annotation

Repetitive elements (REs) across all 17 species were predicted through a combination of evidence-based and *ab initio* methods. For the evidence-based method, we predicted repeats within the target genome using RepeatMasker with the following parameters "-a -nolow -no_is -norna" and RepeatProteinMask with parameters of "-engine ncbi -noLowSimple -pvalue 0.0001" (vopen-4.0.9) (Chen, 2004) based on the Repbase (v24.06) (Jurka, 2000). For the *ab initio* method, we first constructed a *de novo* repeat library of the target genome using RepeatModeler (v2.0) with the parameter "-engine rmbast". Long terminal retrotransposons (LTRs) were identified using both LTR_FINDER_parallel (v1.1) (Ou and Jiang, 2019) with the following parameters "-harvest_out -size 1000000 -time 300 -finder" and LTRharvest v1.0 (Ellinghaus et al., 2008) with the following parameters "-minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20 -seqids yes". Then, the LTRs candidates were further passed to LTR_retriever (v2.8) (Ou and Jiang, 2018) with default parameters to filter out false LTRs, and calculate the LTR Assembly Index (LAI). Finally, the repeat libraries from LTR_retriever and RepeatModeler were merged to complete *de novo* prediction of REs using RepeatMasker with the following parameters "-nolow -no_is -norna". In addition, tandem repeats were predicted by using the Tandem Repeat Finder (TRF v4.09) package (Benson, 1999) with the following parameters "2 7 7 80 10 50 2000 -d -h".

The prediction of protein-coding genes in the *L. philippensis* genome involved the integration of three distinct methods, including *ab initio* gene prediction, homology-based gene

prediction, and RNA-Seq-assisted gene prediction. Before proceeding with protein-coding gene prediction, we soft-masked the assembled *L. philippensis* genome using Bedtools (Quinlan and Hall, 2010) according to the annotated file of TEs. For *ab initio* gene prediction, we employed GenScan (v1.0) (Aggarwal and Ramaswamy, 2002), GlimmerHMM (v3.0.3) (Majoros et al., 2004), Augustus (v3.2.2) (Stanke et al., 2008), and SNAP (v1.0) (Korf, 2004) to predict protein-coding genes. Next, homology-based gene prediction was performed using TBLASTN (Altschul et al., 1990) with a cutoff threshold of $1e^{-5}$, searching against protein sequences from five reference species, including *A. thaliana*, *V. vinifera*, *Solanum lycopersicum*, *S. indicum*. To execute RNA-Seq-assisted gene prediction, the transcriptome assembly was used for gene prediction by comparing it with genomes using the Program to Assemble Spliced Alignments (PASA) (Haas et al., 2003). Finally, a non-redundant gene set was integrated using EvidenceModeler (v1.1.1) (Haas et al., 2008) and updated with PASA. Based on sequence similarity and domain conservation, functional annotations of gene models were predicted by the online EggNOG (v5.0.0) database (Cantalapiedra et al., 2021).

Phylogenetic reconstruction and comparative genomics analysis

The longest protein-coding sequences of *L. philippensis* and the other 16 species were extracted and clustered using OrthoFinder (v2.5.2) (Emms and Kelly, 2019). Subsequently, the protein-coding sequences of single-copy gene were subjected to multiple sequence homology alignment using Mafft (v7.471) (Katoh and Standley, 2013) with the following parameters “-localpair -maxiterate 1000”. Each coding sequence (CDS) was aligned separately according to the corresponding amino acid alignments using PAL2NAL (v14) (Suyama et al., 2006), and then all CDS matrixes were concatenated into a supermatrix. After filtering the poorly aligned regions of integrated CDS alignments using Gblocks (v0.91b) (Castresana, 2000), a maximum likelihood (ML) tree was constructed using IQ-TREE v2.2.0.3 (Kalyaanamoorthy et al., 2017) with the following parameters “-m MFP; -bb 1000; -nt 10” and with the best-fit model (GTR+F+I+G4). Divergence times for single-copy gene supermatrix dataset were estimated based on the ML tree using MCMCTree module from the PAML package with the following parameters “burnin = 50000; nsample = 100000” (Yang, 2007). Two fossil calibration points for divergence time estimation were searched from the TimeTree database (<http://www.timetree.org/>). One is *L. philippensis* versus *V. vinifera* (range: 111.4~123.9 Mya) and another is *L. philippensis* versus *B. alternifolia* (range: 31.5~56.1 Mya). The resulting phylogenetic tree was visualized using FigTree (v1.4.3) (<https://github.com/rambaut/figtree>). The expansion and contraction of gene family in *L. philippensis* were determined using Computational Analysis of Gene Family Evolution (CAFE v5.0) (Mendes et al., 2021) with the following parameters “-k 3 -cores 30”. This process through comparing orthologs groups of itself with other 16 species based on the cluster results of OrthoFinder (v2.5.2) (Emms and Kelly, 2019) and the ultrametric phylogeny generated

from r8s (Mulcahy et al., 2012). Finally, ortholog groups with $P < 0.05$ were considered as gene families undergoing significant expansion or contraction. The correlation between genome size and repeat content was calculated using the “cor.test” function in R 4.2.1 with the Pearson method.

Analyses of whole-genome duplication

The WGD events experienced by *L. philippensis* and the other 16 species were determined by combining the analysis of synonymous substitutions per synonymous site (Ks) and the syntenic analysis that reflects the syntenic depth of intergenomic collinear blocks.

Firstly, syntenic blocks (paralogous genes) within each species were identified using WGDI (v0.6.2) (Sun et al., 2022) with the parameters “-d, -icl, -ks, -bi, -c, -bk”, and then the Ks between collinear genes were calculated by using the Nei-Gojobori approach as implemented in the PAML (v4.9h) package (Yang, 2007). Median Ks values were used to represent each syntenic block, and Ks peak fitting was performed using WGDI with the “-pf” option (Sun et al., 2022). Secondly, the syntenic depths of collinear genes within other species were employed to determine syntenic ratios between different species, confirming their polyploidy levels. To exactly detect the polyploidization levels, we detected the syntenic depth via two methods. One method involved using WGDI (Sun et al., 2022) with the “-bk” option, while the other one utilized JCVI (v1.3.8) with two sets of parameters: “jcv.compara.catalog ortholog; -no_strip_names -cscore=0.99” and “jcv.compara.synteny depth -histogram” (Tang et al., 2008).

Inference of Lamiales ancestral karyotype and analyses of karyotype evolutionary trajectory

We used the ‘Telomere-centric genome repatterning model’ proposed in previous study (Wang et al., 2015; Sun et al., 2022) to construct the LAK and infer its evolutionary trajectory in Lamiales plants. Given the conserved karyotype of *L. philippensis*, its genome was chosen to complete the construction of LAK. The construction process was delineated into three key steps: Step 1 entailed the detection of Whole Genome Duplication (WGD); Step 2 involved the reconstruction of the ancestral karyotype; and Step 3 focused on validating the accuracy of the reconstructed ancestral karyotype. A more detailed description was provided in Supplementary Materials (Note 1).

To analyze the evolutionary history of karyotype among Lamiales species, 13 species from eight families were chosen. Similar to the process of LAK inference, we utilized WGDI (Sun et al., 2022) with the parameters “-d, -icl, -bi, -c, -km, -d” to complete karyotype mapping between different species with LAK. Additionally, the dynamic evolutionary trajectory of LAK and post-LAK following the γ -WGT event was illustrated using Adobe Animate software.

Construction of subgenome and comparative analyses of eight Lamiales species

Eight species, each representing a distinct family and possessing a relatively complete ancestral karyotype, were selected to investigate the traits of karyotype evolutionary trajectory. To precisely build the sub-genomes, two LAK copies in *L. philippensis* (post-LAK1-22) were created to aid in constructing the sub-genome of other species. Similar to the previous reconstruction of LAK, we utilized WGDI with the parameters “-d, -icl, -bi, -c, -km, -ak, -d” to construct the sub-genomes of the eight species. The syntenic relationship between the 16 subgenomes was then visualized using JCVI (Tang et al., 2008). For further characterization of the subgenome, each subgenome was tackled as species, and their corresponding protein-coding sequences were clustered into orthogroups using OrthoFinder (v2.5.2) (Emms and Kelly, 2019). The intersection of different groups was visualized using a website tool at <https://bioinformatics.psb.ugent.be/webtools/Venn/>.

The identification of different modes of gene duplication and the analysis of CYP superfamily

Various gene duplication modes were identified utilizing the “DupGen_finder-unique.pl” module of DupGen_finder (Qiao et al., 2019) with default parameters, and *O. pumila* was set as the reference. We identified CYP genes using HMMER v3.3.2 (Potter et al., 2018) with parameter “-cut_tc”. The Pfam HMM models, namely PF00067 was set as queries for the identification of CYPs. The previously characterized *A. thaliana* CYPs genes was downloaded from <http://p450.kvl.dk/index.shtml> and used as outgroups. To construct the phylogenies for CYPs, the protein sequences were aligned using MAFFT (Katoh and Standley, 2013). The poor alignments were trimmed using trimAl (Capella-Gutiérrez et al., 2009). ML phylogenetic trees were constructed with IQ-TREE (Kalyaanamoorthy et al., 2017) and visualized using iTOL (Letunic and Bork, 2024).

Results

Genome assembly and annotation of *Lindenbergia philippensis*

Through the analysis of 17-kmer frequencies from Illumina short-reads and flow cytometry, the genome size of *L. philippensis* was estimated at approximately 416.78 Mb and 396.66 Mb, respectively, with a heterozygosity rate of 0.706% (Supplementary Tables S1, S2; Supplementary Figures S1, S2). The consistency of genome size estimation was observed between these two methods. A total of 40.13 Gb (101×) of raw ONT long-reads were utilized for the initial assembly of contigs using NextDenovo (v2.5) (<https://github.com/Nextomics/NextDenovo>) (Supplementary Table S3).

After two rounds of polish of the 80.49 Gb (202×) Illumina short-reads using Nextpolish v1.2.1 (<https://github.com/Nextomics/NextPolish>), we obtained 949 final contigs with a total size of 406.79 Mb and a N50 of 1.79 Mb (Supplementary Tables S3, S4). Subsequently, the polished contigs were clustered and ordered using 131.51 Gb (331×) Hi-C data through Juicer (Durand et al., 2016) and 3D-DNA (Dudchenko et al., 2017), resulting in the successful construction of 16 pseudo-chromosomes with a scaffold N50 of 23.51 Mb, covering approximately 96.55% of the final assembled sequences (393.39 Mb/407.46 Mb) (Figure 1A; Supplementary Figure S3).

The final assembled genome size was nearly close to the size estimated by the flow cytometry and the 17 kmer frequency distribution (Supplementary Figure S1; Supplementary Table S4). Furthermore, mapping 536,633,198 Illumina reads to the final assembly resulted in a mapping rate of 99.15% and a coverage rate of 95.05% (Supplementary Table S5). The completeness of genome assembly is 98.6% of BUSCO genes based on the embryophyta_10 dataset (Seppey et al., 2019), which was comparable with 14 genomes of Lamiales species (Figure 1B; Supplementary Table S6). *Lindenbergia philippensis* genome had a high Long-terminal repeat (LTR) Assembly Index (LAI) score of 12.9 (Figure 1B), meeting the “reference standard” (LAI value > 10) of genome assembly proposed by Finn et al. (2011).

Based on homologous and *de novo* prediction, 250.16 Mb of repetitive elements (REs) were identified in the *L. philippensis* genome, constituting 61.40% of the assembly genome. These elements included LTRs (39.68%), DNA transposons (6.63%), LINEs (0.86%), SINEs (0.02%), and unclassified sequences (15.78%) (Supplementary Table S8). After masking the REs, 25,693 protein-coding genes were identified by combining *de novo*, homology-based, and RNA-Seq-based predictions. On average, each predicted gene had an average length of 3,800 bp and contained five exons with an average length of 232 bp (Supplementary Table S9). Approximately 94.89% of protein-coding genes were functionally annotated by existing databases (Supplementary Table S10).

Comparative and evolutionary genomics of *Lindenbergia philippensis* and its relatives

To investigate genomic characteristics of *L. philippensis* and its relatives, comparative genomic analyses were performed on 15 representative genomes from 12 families of Lamiales and two outgroups, *V. vinifera* and *O. pumila* (Figure 1E, Supplementary Table S7). The annotation and comparison of their REs revealed that the repeat size was widely distributed in these 17 genomes, varying from 88.8 Mb to 1,242.6 Mb, with *O. cumana* exhibiting the highest repeat content (Figure 1B; Supplementary Table S8). Meanwhile, correlation analysis showed that the genome size was positively correlated with the repeat contents ($R = 0.97$, $P < 0.05$), which was consistent with previous studies (Figure 1C) (Shao et al., 2019; de Lima and Ruiz-Ruano, 2022).

By employing OrthoFinder (v2.5.2) (Emms and Kelly, 2019) to cluster orthologs, a total of 576,537 genes from 17 genomes were

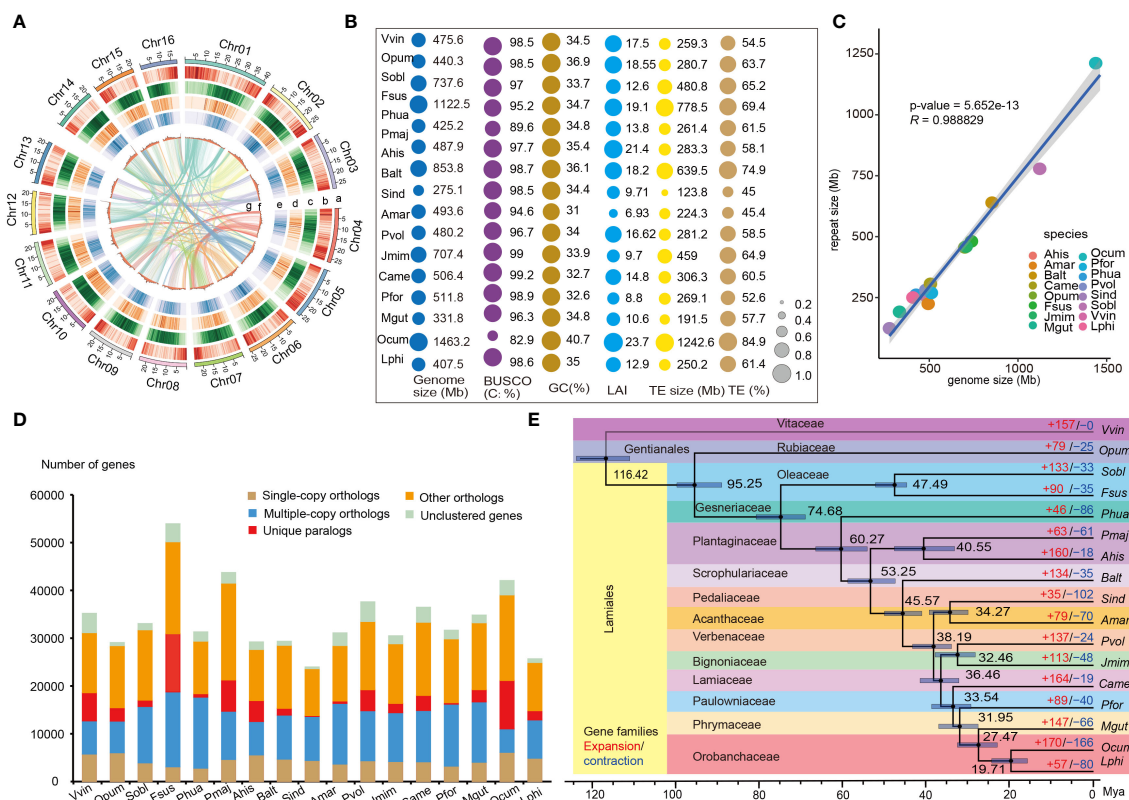


FIGURE 1

Genomic features and comparative analysis of *L. philippensis* with other 16 species. (A) The genomic features are arranged in the order of pseudo-chromosomes (scale is in Mb), gene density, repeat density, LTR/Gypsy, LTR/Copia, GC contents, and syntenic blocks from outside to inside in 300 kb intervals across the 16 pseudo-chromosomes. (B) Comparative analysis of genomic quality index in *L. philippensis* (Lphi) with other 16 species, *P. huaijiensis* (Phua), *F. suspense* (Fsus), *S. oblate* (Sobl), *A. hispanicum* (Ahis), *P. major* (Pmaj), *C. alternifolia* (Balt), *S. indicum* (Sind), *A. marina* (Amar), *P. volubilis* (Pvol), *(J) mimosifolia* (Jmim), *C. americana* (Came), *P. fortune* (Pfor), *M. guttatus* (Mgut), *O. cumana* (Ocum), *V. vinifera* (Vvin) and *O. pumila* (Opum). The size of the colored round shapes represents the number or proportions of all indexes in each species. (C) Analysis of the correlation between genome size and RE content among 17 species. (D) Distribution of single- and multiple-copy, and other orthologs, unique paralogs, and unclustered orthologs per species from orthogroup clustering by OrthoFinder (v2.5.2) (Emms and Kelly, 2019). (E) Phylogenetic tree inferred from single-copy orthologs among selected species. Black numbers in each node denote the divergence time of each clade (Mya), and gray bars are 95% confidence intervals for the time of divergence between different clades. The red and the blue numbers at the terminal branches show the expansion (red) and contraction (blue) of gene families for each species.

classified into 540,506 orthologs groups and 36,031 unclustered genes. Among them, 7,775 groups were shared by all 17 species, including 326 single-copy orthologs groups (Figure 1D; Supplementary Table S11). *Lindenbergia philippensis* possessed 1,934 species-specific genes, including 289 orthologs genes and 628 unclustered genes (Figure 1D; Supplementary Table S12). The biological processes of species-specific genes were mainly distributed in 'host cellular response', 'metabolic process' and 'biosynthetic process' (Supplementary Figure S4), suggesting the evolution of key enzyme genes associated with metabolite synthesis and pathways for environmental adaptation in *L. philippensis*. The phylogenetic tree constructed using 326 conserved single-copy genes from 17 genomes using the maximum-likelihood method showed that *L. philippensis* was sister to parasitic species in Orobanchaceae, aligning with prior research (Figure 1E) (Mutuku et al., 2021). Divergence time estimation showed that the divergence between *L. philippensis* and *O. cumana* occurred at ~19.71 million years ago (Mya), and the Lamiales diverged from the Gentianales at ~95.25 Mya (Figure 1E). Expansion or contraction of gene families

is often associated with adaptive divergence in closely related species (Cheng et al., 2017). Therefore, we investigated changes in gene families using the estimated phylogeny to capture key genomic information associated with *L. philippensis* adaptability. Compared to related species, a total of 57 gene families (including 496 genes) and 80 gene families (including 90 genes) exhibited significant expansion and contraction in the *L. philippensis* genome, respectively ($P < 0.05$) (Figure 1E). Interestingly, the expanded genes were primarily enriched in many secondary metabolite biosynthetic pathways (e.g., flavonoid biosynthesis and metabolic process, glucan metabolic process and cellulose biosynthetic process), suggesting that *L. philippensis* produces some active substances such as phenols (Supplementary Figure S5).

Polyploidization history of Lamiales species

To unveil the ancient polyploidization history of Lamiales species, we examined the distribution of substitutions per

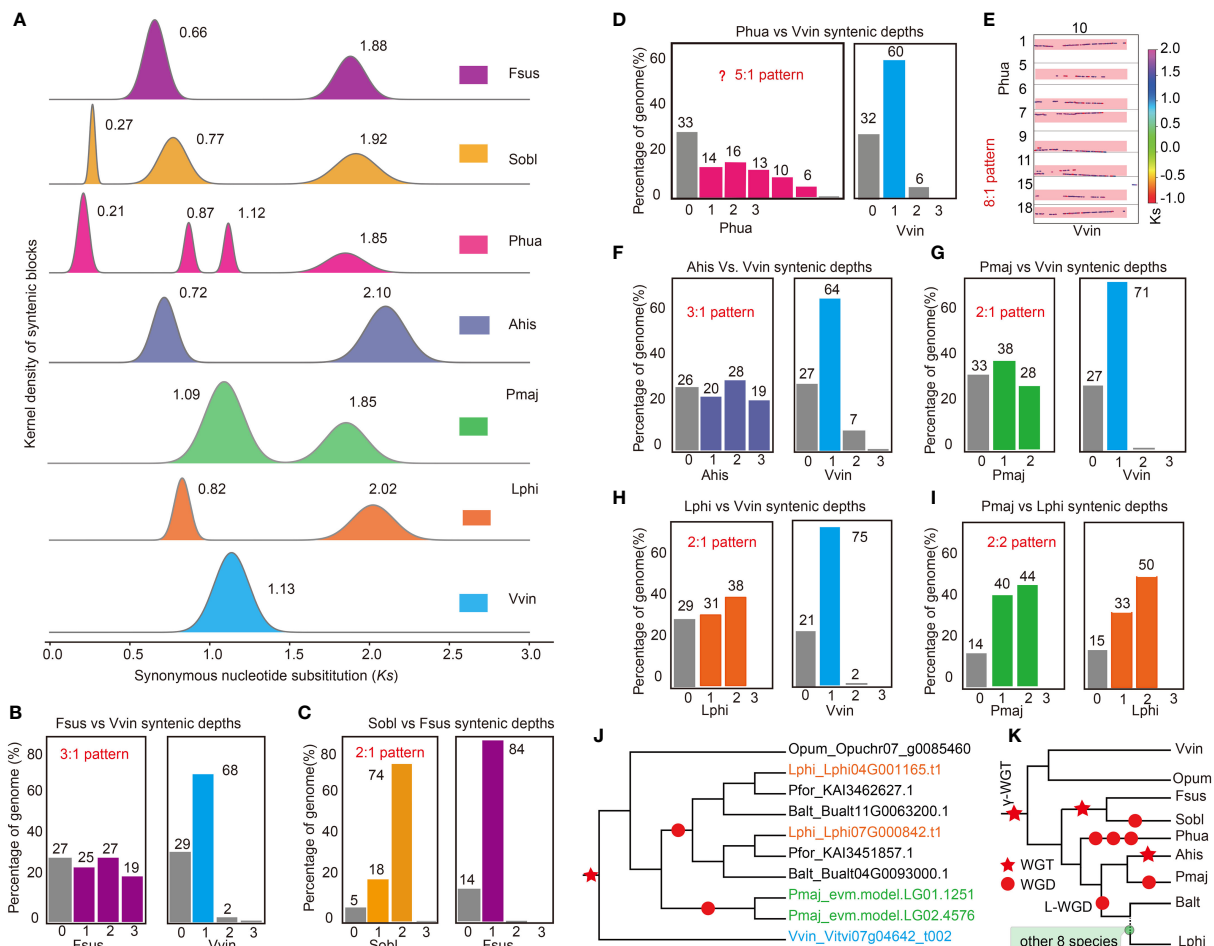


FIGURE 2
Inference of polyploidization histories in the genomes of the studied Lamiales species. (A) The synonymous substitution (Ks) distributions of gene pairs in syntenic blocks among compared genomes. (B) The ratio of orthologous genes between *F. suspense* (Fsus) and *V. vinifera* (Vvin). (C) The ratio of orthologous genes between *S. oblate* (Sobl) and *F. suspense* (Fsus). (D) The ratio of orthologous genes between *P. huaijiensis* (Phua) and *Vvin*. (E) The syntenic depth of homologues blocks between *Phua* and *Vvin*. (F) Ratio of orthologous genes between *A. hispanicum* (Ahis) and *Vvin*. (G) The ratio of orthologous genes between *P. major* (Pmaj) and *Vvin*. (H) The ratio of orthologous genes between *L. philippensis* (Lphi) and *Vvin*. (I) The ratio of orthologous genes between *P. major* (Pmaj) and *Lphi*. (J) The phylogenetic tree of ten orthologous genes, derived from four Lamiales species, *Vvin* and *O. pumila* (Opum). (K) Overview of WGD events in those 15 Lamiales species. Polyploidization events are indicated by red pentagram (triploidization, WGT) and red round shape (diploidization, WGD).

synonymous site (Ks) of intra-genomic collinear blocks in the 15 Lamiales species. Two to four separate peaks were detected in the Ks distribution for species-specific paralogous pairings in those species (Figure 2A, Supplementary Figure S6), indicating that at least one round of WGD events occurred in this lineage following the γ -WGT event. For example, four obvious Ks peaks were observed in *P. huaijiensis*, reflecting a younger WGD event at Ks 0.21, two distinct WGD events at Ks 0.87 and 1.12 respectively, and γ -WGT event at Ks 1.85. In *S. oblate*, three Ks peaks indicated a younger WGD event at Ks 0.27, a WGD event at Ks 0.77, and the γ -WGT event at Ks 1.92 (Figure 2A). Other species such as *F. suspensa*, *A. hispanicum*, *P. major*, *B. alternifolia*, *S. indicum*, *P. volubilis*, *P. fortunei*, *J. mimosifolia*, *C. americana*, *M. guttatus*, *O. cumana* and *L. philippensis* exhibited two peaks (Figure 2A; Supplementary Figure S6). The first peak indicated a recent WGD event, while the second peak corresponded to the γ -WGT event. *Vitis vinifera* and *O. pumila* displayed only a single peak representing γ -WGT

event (Figure 2A; Supplementary Figure S6). The distribution of Ks peaks showed differences among species (Figure 2A; Supplementary Figure S6), which was usually caused by evolutionary rate variations in habitat divergence (Sensalari et al., 2022). For example, besides *V. vinifera*, *P. fortunei* exhibited the lowest Ks value in γ -WGT event (Supplementary Figure S6), indicating it may have a lower evolutionary rate than other species.

To determine the polyploidization level of Lamiales species after the γ -WGT event, their ratio of orthologous genes with *V. vinifera* was examined with precision. Generally, a species experienced the WGD event will have a corresponding orthologous gene ratio with another species, which only retained their common ancestral karyotype. For instance, Hoang et al. (2023) demonstrated that *Cleome violacea* did not undergo Gg- α (diploidization) event after divergence from a shared ancestor with *Gynandropsis gynandra*, which had experienced a diploidization event. Consequently, *C. violacea* exhibited a 1:2 orthologous gene ratio with *G. gynandra*.

Through comparative analysis of genome syntenic blocks, we have successfully determined the level of polyploidization in those 17 species following γ -WGT event. *Ophiorrhiza pumila* exhibited a 1:1 orthologous gene ratio with *V. vinifera* (Supplementary Figure S7), suggesting that it underwent only the shared γ -WGT event and did not experience additional WGD events after diverging from their common ancestor. *Forsythia suspensa* had a 3:1 orthologous gene ratio with *V. vinifera* (Figure 2B; Supplementary Figure S8), indicating it experienced a triploidization at Ks peak 0.66. *Syringa oblata* had a 6:1 orthologous gene ratio with *V. vinifera* and a 2:1 orthologous gene ratio with *F. suspensa*, respectively (Figure 2C; Supplementary Figures S9, S10), indicating it experienced a common triploidization with *F. suspensa* at Ks peak 0.77 and an independent diploidization event at Ks peak 0.27 (Figure 2A). The two rounds of WGD events were also proved by previous studies (Julca et al., 2018; Feng et al., 2020). *Primulina huaijiensis* showed an 8:1 orthologous gene ratio with *V. vinifera* (Figures 2D, E; Supplementary Figure S11). Therefore, the orthologous gene ratio between *P. huaijiensis* and *V. vinifera* could be explained as $2 \times 2 \times 2:1$ according to the Ks distribution, corresponding to three rounds of diploidization events rather than two rounds of diploidization events reported in a previous study (Feng et al., 2020). *Antirrhinum hispanicum* had a 3:1 orthologous gene ratio with *V. vinifera* (Figure 2F; Supplementary Figure S12), indicating that it underwent a triploidization event at Ks peak 0.72 (Figure 2A), which was consistent with previous results (Zhu et al., 2023b). *Avicennia marina* had a 4:1 orthologous ratio with *V. vinifera* and a 2:1 orthologous gene ratio with *L. philippensis* (Supplementary Figures S13; S33), respectively, indicating it experienced a common diploidization with *L. philippensis* at Ks peak 0.77 and an independent diploidization event at Ks peak 0.27 (Figure 2A).

In addition, other ten species, including *P. major*, *B. alternifolia*, *S. indicum*, *P. volubilis*, *P. fortunei*, *J. mimosifolia*, *C. americana*, *M. guttatus*, *O. cumana* and *L. philippensis*, had a 2:1 orthologous gene ratio with *V. vinifera* (Figures 2G, H; Supplementary Figures S14, S23). This suggests that they could have experienced a common diploidization event after the γ -WGT event, corresponding to *L_event* revealed by previous results (Julca et al., 2018; Feng et al., 2020). To better determine whether the ten species underwent a common diploidization event, we examined the inter-genomic collinearity relationships among orthologous genes, using *L. philippensis* as the reference. Except for *P. major*, which exhibited a 2:2 orthologous gene ratio with *L. philippensis* (Figure 2I; Supplementary Figure S24), the remaining eight species displayed a 2:1 orthologous gene ratio with *L. philippensis* (Supplementary Figures S25–S32). This suggests that *P. major* may have undergone a diploidization event independently, while the remaining nine species shared a common diploidization event after the γ -WGT event. Furthermore, phylogenetic analyses of orthologous genes derived from four paired subgenomes and using *O. pumila* and *V. vinifera* as outgroups, further corroborating this hypothesis (Figure 2J).

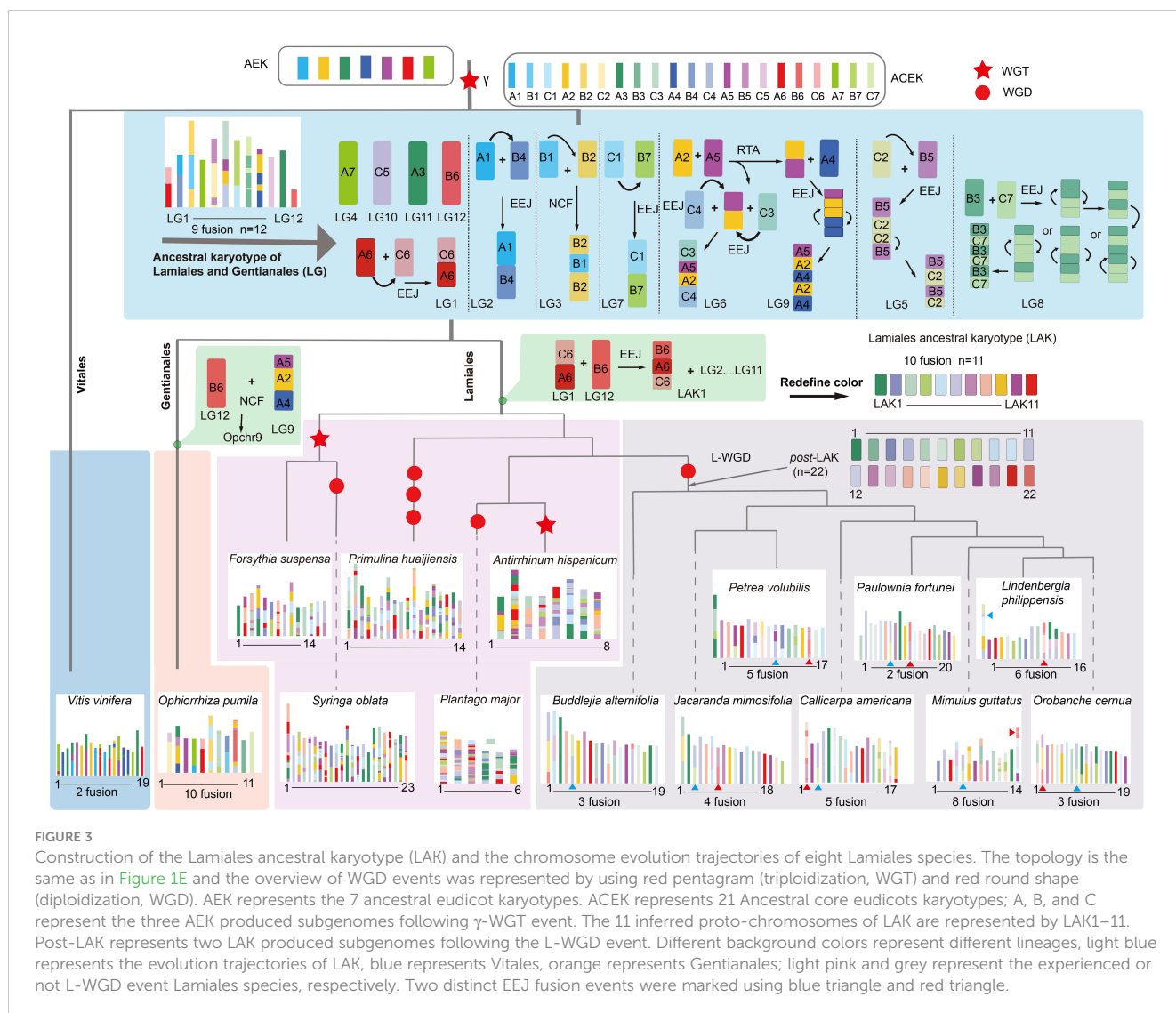
In summary, after the γ -WGT event, Lamiales species underwent multiple WGD events based on Ks and syntenic analyses. *Syringa oblata* and *F. suspensa* underwent a shared triploidization event, and *S. oblata* subsequently underwent an independent diploidization event, in line with the previous

finding (Julca et al., 2018; Feng et al., 2020). *Primulina huaijiensis* experienced three rounds of diploidization events. *Antirrhinum hispanicum* experienced a triploidization event, while *P. major* underwent a diploidization event. Integrating phylogenetic and syntenic analyses, we found that the remaining ten species from nine families, including *B. alternifolia* (Scrophulariaceae), *S. indicum* (Pedaliaceae), *A. marina* (Acanthaceae), *P. volubilis* (Verbenaceae), *J. mimosifolia* (Bignoniaceae), *C. americana* (Lamiaceae), *P. fortunei* (Paulowniaceae), *M. guttatus* (Phrymaceae), *O. cumana*, and *L. philippensis* (Orobanchaceae), underwent a shared diploidization event, known as L-WGD event (Figure 2K).

Construction of Lamiales ancestral karyotype and analyses of karyotype evolutionary trajectories

After polyploidization, substantial karyotype changes frequently occur in many plant genomes. These changes can alter the basic chromosome number and trigger species diversification. In Lamiales, the chromosome numbers of 15 selected species range from $2n=12$ to $2n=64$ (Supplementary Table S7). These variations are primarily caused by karyotype changes. To uncover the karyotype evolutionary trajectories, the *L. philippensis* genome was used to reconstruct the Lamiales ancestral karyotype (LAK).

To comprehensively delineate the karyotype evolutionary trajectory of the LAK in Lamiales species, we defined the 21 proto-chromosomes of the Ancestral Core Eudicot Karyotypes (ACEK), derived from the triplication of seven ancestral eudicot karyotypes (AEK), as A1-7, B1-7, and C1-7. As a result, a putative LAK was constructed consisting of 11 proto-chromosomes (LAK1–LAK11), which shared the same base chromosomal number with the sister clade species such as *O. pumila* ($2n=22$), *Morinda officinalis* ($2n=22$), and *Leptodermis oblonga* ($2n=22$) from the Gentianales order. This suggests a possible common ancestral karyotype between Lamiales and Gentianales. To validate this hypothesis, we generated a dot plot by comparing the LAK and *O. pumila* (Rubiaceae family) with ACEK. Rubiaceae, positioned at the root of Gentianales, exhibits a higher likelihood of sharing the same karyotype with LAK among its species. The dot plot analysis indicated that nine chromosomes of the LAK and *O. pumila* exhibited a one-to-one correspondence in their collinearity relationship (Supplementary Figures S34–S36). The primary distinction between them lies in the rearrangement of proto-chromosome B6 (Figure 3). Following the methodology used in constructing the LAK, we constructed a hypothetical common ancestral karyotype for Lamiales and Gentianales orders, labeled as LG1–LG12 (Figure 3; Supplementary Figure S37). In summary, LAK evolved into 11 proto-chromosomes through a series of chromosomal rearrangements, including nine end-to-end joining (EEJ), two nested chromosome fusions (NCF), and ten reciprocal translocations of chromosome arms (RTA). For example, the formation of proto-chromosomes LAK1 was mainly explained by the fusion of A6 and C6 initially with the EEJ pattern and then further fused with B6 through the EEJ pattern (Figure 3). Similarly,



the evolutionary trajectory of the other ten proto-chromosomes of LAK was inferred in Figure 3.

Based on previous results, eight Lamiales species, including *B. alternifolia*, *P. volubilis*, *P. fortunei*, *J. mimosifolia*, *C. americana*, *M. guttatus*, *O. cumana* and *L. philippensis*, have been identified as sharing the L-WGD event. This makes them ideal candidates for exploring the evolutionary characteristics of the LAK. Following the L-WGD event, the LAK underwent duplication, resulting in the formation of 22 proto-chromosomes (post-LAK). To investigate the evolutionary characteristics of the post-LAK in these species, two sets of LAK generated from *L. philippensis* were used to represent post-LAK karyotype and labeled as post-LAK1 to post-LAK22. Two distinct EEJ fusion events were identified by analyzing the dot plot comparing these eight species with the post-LAK (Figure 3). The first EEJ fusion event, which involved post-LAK4 and post-LAK8, occurred in all eight species. In contrast, the second EEJ fusion event, involving post-LAK20 and post-LAK22, was only present in seven of these species, with *B. alternifolia* being the sole exception (Figure 3). Subsequently, the eight species separated and evolved with different chromosome evolutionary trajectories.

Given the non-dysploid chromosomal changes were prevalent, we focused primarily on depicting the dysploid chromosomal rearrangements in these species. In summary, *B. alternifolia* genome experienced three chromosomal fusions, consisting of two EEJ fusions and one NCF fusion, leading to the current chromosome number $n=19$ (Figure 3; Supplementary Figure S38); *J. mimosifolia* genome experienced two EEJ fusions and two NCF fusions to form the current chromosome number $n=18$ (Figure 3; Supplementary Figure S39); the *P. volubilis* genome experienced five chromosomal fusions, composing of four EEJ and one NCF, resulting in the current chromosome number $n=17$ (Figure 3; Supplementary Figure S40); *C. americana* genome experienced three EEJ fusions, one NCF fusion and one EEJ or NCF fusion, leading to the current chromosome number $n=17$ (Figure 3; Supplementary Figure S41); *P. fortunei* genome experienced the fewest karyotype change events to form the current chromosome number $n=20$, with just two EEJ fusions and no further karyotype evolutionary events (Figure 3; Supplementary Figure S42); *M. guttatus* genome experienced three EEJ fusions and five NCF fusions to form the current chromosome number $n=14$ (Figure 3;

Supplementary Figure S43); *O. cumana* genome experienced two EEJ fusions and one NCF fusion to form the current chromosome number $n=19$ (Figure 3; Supplementary Figure S44); *L. philippensis* genome experienced two EEJ fusions and four NCF fusions to form the current chromosome number $n=16$ (Figure 3; Supplementary Figure S45).

Comparative analyses of subgenomes in eight Lamiales species

Following polyploidization, most duplicated genes would experience drastic changes due to the sensitivity of dosage balance (Li et al., 2016). To elucidate the fractionation characteristics of duplicated genes in Lamiales, two sets of post-LAK subgenomes were initially classified as least fractionated (LF, 22A) and most fractionated (MF, 22B) based on their gene counts. Subsequently, 16 subgenomes were constructed using the WGDI, and their grouping was determined based on the collinear relationship with post-LAK. The one-to-one collinear correspondence of these subgenomes with post-LAK confirmed the reliability of these subgenomes (Figure 4A; Supplementary Figures S46–S52), making them suitable for further

research. Like the post-LAK, all subgenomes exhibited subgenome dominance. For example, 22A subgenomes (with 14,306 – 24,133 genes) had more gene counts than 22B subgenomes (with 9,364 – 16,727 genes) among these 16 subgenomes (Supplementary Table S13). The BUSCO analyses also showed that eight 22A subgenomes had over 50% complete BUSCO genes from the embryophyta_10 dataset, whereas the completeness level in the eight 22B subgenomes was below the threshold of 50% (Figure 4B). This phenomenon suggested that these eight species exhibit consistently biased preservation and display a dominance within their respective subgenomes.

To uncover the fractionation pattern of the subgenomes, we utilized OrthoFinder (v2.5.2) (Emms and Kelly, 2019) to group their protein-coding genes into orthogroups, with *V. vinifera* and *O. pumila* as the reference. Stringent criteria were applied to choose representative orthogroups, necessitating orthogroups with a minimum of eight distinct subgenomes, encompassing *V. vinifera* and *O. pumila*. In total, 10,083 orthogroups were selected to comprise the core set of orthogroups (CSOs) for our further analyses. Based on the observed number of gene copies in each CSO, those CSOs were categorized into four distinct types: ‘Absent’ (no gene copies present), ‘Single Copy’ (one gene copy), ‘Two

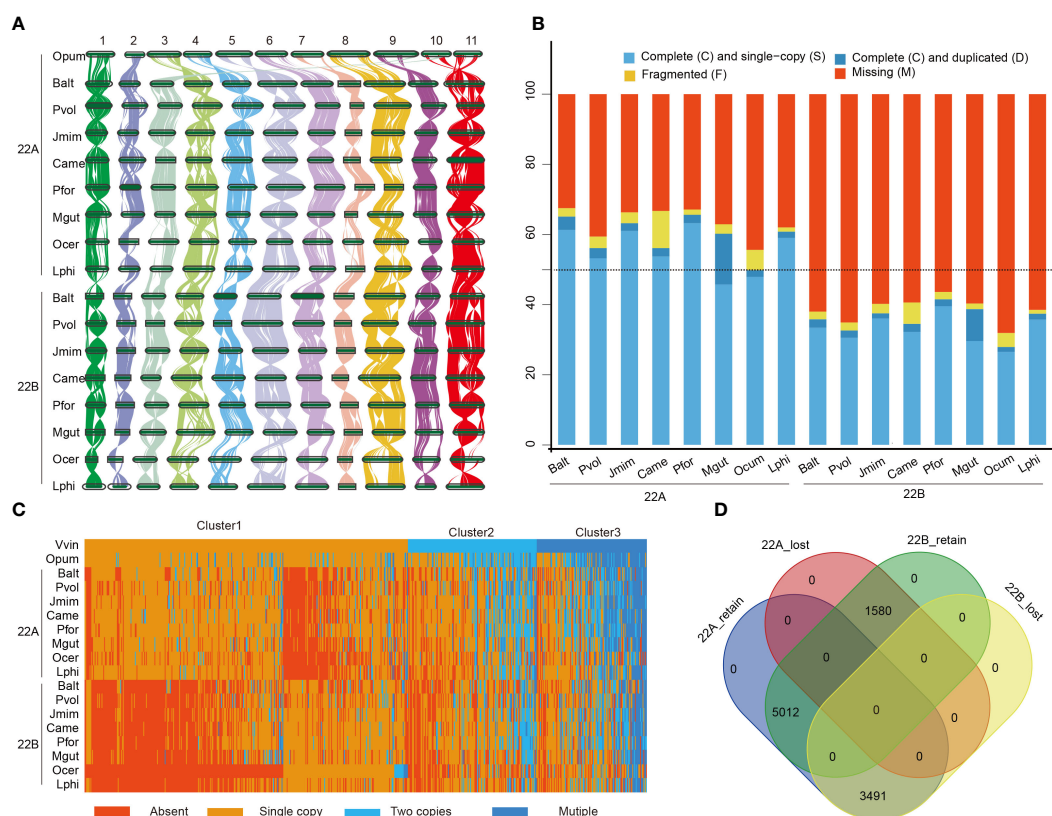


FIGURE 4

Comparative analysis of subgenomes of 8 species in the Lamiales. (A) The synteny plot across *O. pumila* (Opum) genome and sixteen subgenomes, *B. alternifolia* (Balt), *P. volubilis* (Pvol), *J. mimosifolia* (Jmim), *C. americana* (Came), *P. fortune* (Pfor), *M. guttatus* (Mgut), *O. cumana* (Ocum) and *L. philippensis* (Lphi); 22A represents least fractionated subgenome and 22B represents most fractionated subgenome. (B) Assessment of Benchmarking Universal Single-Copy Orthologs (BUSCOs) of those sixteen subgenomes with embryophyta_10 (1614) databases. (C) Heat map of the clustered copy-number profile matrix in Opum, *V. vinifera* (Vvin), and sixteen subgenomes. Core gene families could be partitioned into four based on the clustering of the copy-number profile data. Rows represent species and columns represent the 10,083 CSOs. Gene families are sorted according to the three different clusters of Vvin. (D) Venn diagram showing the distribution of the retained and lost CSO sets.

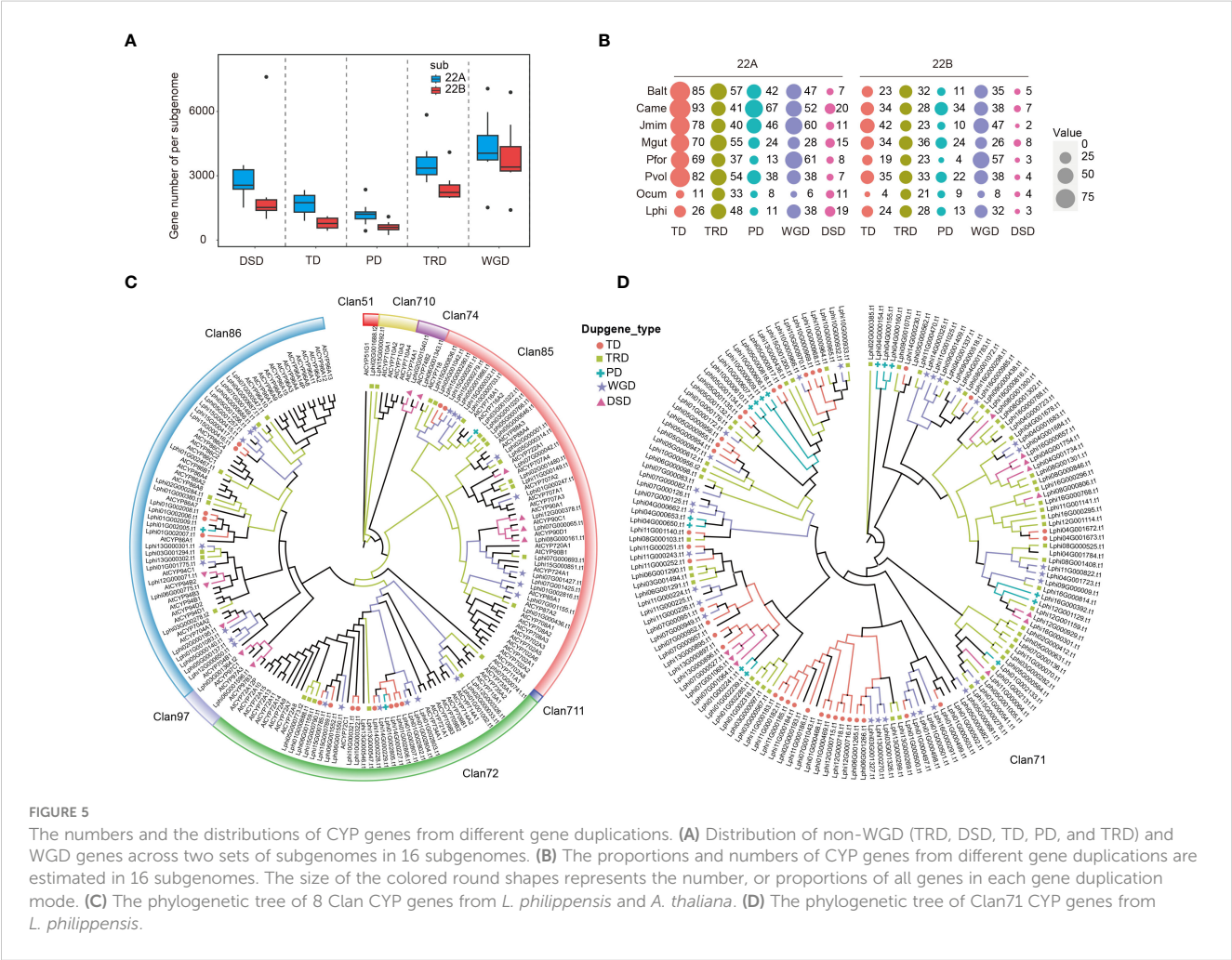
Copies' (exactly two gene copies), and 'Multiple' (more than two gene copies). Furthermore, according to the gene number in *V. vinifera*, those CSOs were organized into three clusters (cluster 1, cluster 2, and cluster 3) (Figure 4C). In cluster 1, CSOs mainly consisted of absent or singleton genes in 16 subgenomes (Supplementary Figure S53). In cluster 2, CSOs are composed of either absent or present genes in single or two-copy forms. In cluster 3, CSOs mainly consisted of orthogroups that are single, two, or multiple copies. In all three clusters, the number of absent CSOs in the 22B subgenome was significantly greater than that in the 22A subgenome ($P < 0.05$). Conversely, the remaining three types showed the opposite trend, except for the single copy gene in cluster 3 (Supplementary Figure S53). Therefore, we speculated that the dominance of the subgenome could primarily originate from a higher frequency of loss and a lower rate of retention. Besides, our results also indicated that the distribution of CSOs across the 22A and 22B subgenome had a nested complementary profile, particularly evident in cluster 1.

We further defined the CSOs present in over 50% of the subgenome as retained CSO sets, while those not maintained are referred to as lost CSO sets. Based on these criteria, slightly over half of the CSO sets (5,071/10,083) displayed a complementary distribution across the two subgenomes, corroborating earlier

findings presented in the heat map analysis. Specifically, 3,491 CSOs were conserved exclusively in the 22A subgenome sets, 1,580 CSOs were solely retained in the 22B subgenome sets, and 5,012 CSOs were present in both 22A and 22B subgenomes (Figure 4D).

Different modes of gene duplications driving the dominance of subgenome

In addition to WGD events, gene duplication is also a crucial process in expanding the gene family (Fajardo et al., 2023). To determine if gene duplication caused the dominance of subgenome, we conducted statistical analyses on non-WGD (Dispersed duplication, DSD, Tandem duplicate, TD; Proximal duplication, PD; and Transposed duplication, TRD) and WGD genes, as well as on unduplicated genes (UD) within those subgenomes. Our results indicated that all the modes of gene duplications were higher in 22A subgenome sets than in 22B subgenome sets (Figure 5A). Cytochrome P450s (CYPs) form the largest enzyme family in plants, representing around 1% of protein-coding genes in various flowering plants (Liu et al., 2023). They can be ideal candidates to study different modes of gene duplications. The distribution of CYP genes in the various



modes of gene duplications showed more copies in the 22A subgenome sets than in the 22B subgenome sets across the 16 subgenomes (Figure 5B). Interestingly, there are more CYP genes in TD genes than in WGD genes, and the number of CYP genes in TRD and PD was also similar to that in WGD genes, despite their lower total gene count compared to WGD genes (Figure 5B). This observation suggests that, besides WGD events, the non-WGD genes also play a crucial role in the expansion of gene families. To detail the influence of the gene duplication on gene family expansion, the phylogenetic tree of CYP genes in *L. philippensis* genome were constructed and with *Arabidopsis thaliana* as references. In total, 242 CYP genes in *L. philippensis* genome were cluster into 9 subfamilies according to the result of Williams et al. (2000) (<http://p450.kvl.dk/p450.shtml>) (Figures 5C, D). Within the phylogenetic tree, the gene duplication modes are distinct among major subfamilies like Clan71, Clan72, Clan85, and Clan86. The remaining subfamilies only exhibit one type of gene duplication mode. Clan71, as the largest subfamily in the CYP superfamily, contains more gene duplication copies of various modes than other subfamilies (Figure 5D).

Discussion

Genome assembly of *Lindenbergia philippensis* provides an important genomic resource

Lindenbergia philippensis belongs to the tribe Lindenbergieae, besides the tribe Rehmannieae, and it is the closest autotrophic sister clade to all parasitic plant lineages in the family Orobanchaceae (Mutuku et al., 2021; Jiang et al., 2022; Xu et al., 2022) (Figure 1E). Here, the *L. philippensis* genome was achieved by combining Illumina paired-end sequencing data, Oxford Nanopore data and Hi-C data. The new genome assembly size was 407.46 Mb, close to the estimated size of 396.66 Mb via flow cytometry and 17-kmer frequency estimation (Supplementary Tables S2, S4). The completeness of the genome assembly was comparable with 15 species in Lamiales (Supplementary Table S6). Therefore, the assembly of *L. philippensis* genome had good quality, making it suitable for further analyses. Additionally, the anchored 16 pseudo-chromosomes had good intra-genomic collinear blocks (Note 1), which makes it the high-quality reference genome to deduce the karyotype evolutionary trajectory among relative species. These results provide important genomic resources for further genome study on *L. philippensis* as well as Orobanchaceae in the future.

Combining Ks and syntenic depth analyses reconstruct the accurate evolutionary history of polyploidization and WGD events

Polyploidization, or WGD events, have been identified as a critical mechanism in facilitating species evolution and diversification across a vast majority of plant lineages (Zhang et al., 2019; Clo, 2022). Additionally, the profound influence of

WGD events goes beyond its initial occurrence, and could primarily serve as a catalyst to drive a subsequent PPD process (Soltis and Soltis, 2016; Zhang et al., 2019). However, the PPD process has negative effect on the identification of WGD events and the determination of polyploidization levels.

Currently, although an increasing number of WGD events are being reported through Ks or 4Dtv analyses, syntenic depth analyses, or a combination of these methods, some WGD events are inaccurately determined due to low-quality and limited genomic data and analytical method constraints. For instance, Feng et al. (2020) used Ks analysis to reveal a WGD (the *L* event) present in almost all Lamiales except the lineage of Oleaceae, which conflicted with the results of Zhu et al. (2023b). Zhu and his colleagues substantiated that the Plantaginaceae underwent a distinct WGD event, diverging from the shared *L* event (Feng et al., 2020; Zhu et al., 2023b). This independent WGD event was confirmed in this study, as well as a recent research by Huang et al. (2023). By combining Ks and inter-species syntenic depth analyses, we validated that *P. huaijiensis* experienced three diploidization events following the γ -WGT event, rather than two WGD events in the previous report (Feng et al., 2020). This discrepancy primarily derived from that Feng et al. (2020) relied on the solely Ks analysis to survey the WGD event, without integrating syntenic depth comparisons across different species. Additionally, two separate Ks values (0.87 and 1.12) (Figure 2A) suggested that *P. huaijiensis* underwent two WGD events within a relatively close timeframe. Consequently, these two WGD events could easily be overlooked and misinterpreted as a single event.

WGDI (Sun et al., 2022) and JCVI (Tang et al., 2008) are both popular software options for analyzing WGD events through syntenic depth analysis, but WGDI has advantages over JCVI in distinguishing the level of polyploidization. For example, *O. pumila* and *V. vinifera* had been shown to share the γ -WGT event, the syntenic depths or orthologous gene ratio between them should theoretically be 1:1, ignoring the non-WGD effects, whereas their syntenic depths were determined at 2:2 in the research of Rai et al. (2021) by using JCVI, which cannot identify whether they shared the WGD or not. In our study, the 1:1 orthologous gene ratio of *O. pumila* and *V. vinifera* was validated using WGDI and confirmed that they shared the γ -WGT event, which was aligned with the previous results (Wang et al., 2022c). Besides, the orthologous gene ratio of *P. huaijiensis* compared to *V. vinifera* was showed to be 6:1, corresponding to its three diploidization events. However, their orthologous gene ratio was 5:1 using JCVI, conflicting with its polyploidization history.

Overall, it is imprudent to crudely estimate polyploidization events based solely on the Ks distribution or syntenic depth analysis. While the analysis of Ks can indicate the occurrence of WGD events, it is challenging to clearly distinguish the polyploidization histories. Essentially, Ks analysis only reveals whether the species underwent WGD events, making it hard to ascertain whether the WGD event led to diploidization, triploidization, or other forms of polyploidization. This and previous studies have revealed some misunderstandings regarding the evolutionary history of WGD events, such as the genomic researches of *C. americana* (Hamilton et al., 2020), watermelon (Guo et al., 2013), black pepper (Hu et al., 2019), Olive (Ren et al., 2018), and *Prunus mongolica* (Zhu et al., 2023a). These mistakes

significantly increase the chance of misinterpreting the evolutionary history of these events, hindering our comprehensive understandings of the functional evolution of subgenomes, gene families, pathways, and genomic structures. Integrating genomic collinearity analysis with Ks information provides a more accurate and effective method for inferring polyploidization events, as supported by our findings in this study and previous studies (Kong et al., 2023; Sun et al., 2024). Based on this theoretical framework, Sun et al. (2022) have developed an integrated tool WGDl that combines functions for detecting WGD events, analyzing karyotype evolution, and constructing ancestral karyotypes, among other functions, providing an effective and more accurate method for the WGD events analyses. Using this tool, WGD events of 15 species in Lamiales were corrected and validated, providing significant insights for the analysis of WGD events. Moreover, the L-WGD shared by most Lamiales species was validated by combing the Ks and syntenic depth analyses.

Construction and evolutionary trajectory of ancestral karyotypes in Lamiales

The identification and construction of ancestral karyotypes play a crucial role in confirming the phylogenetic positions of species and elucidating the impact of various polyploidy events on species diversity and evolution (Murat et al., 2017; Kong et al., 2023). The recursive dysploid or non-dysploid changes have reshuffled the ancestral karyotypes of Lamiales, complicating the clear interpretation of polyploidization events (Ren et al., 2018; Feng et al., 2020). In this study, *L. philippensis* was used to construct the LAK, following the theoretical framework that suggested by Sun et al. (2022), consisting of 11 proto-chromosomes. The two complete copies of the paleogenome within the *P. fortunei* genome validated its reliability (Supplementary Figure S42).

The evolutionary path analyses of LAK and post-LAK showed that the base number deduction of chromosomes was caused by fusions (Wang et al., 2022c; Feng et al., 2024). This suggested that descending dysploidy may play a major role in karyotype evolution after WGD events, consistent with previous studies indicating that the chromosomal evolution in land plants is mostly characterized by descending dysploidy (Carta et al., 2020; Wang et al., 2022c; Kong et al., 2023). Two distinct EEJ fusion events were detected in those eight species, the first fusion shared by all studied species, while the second fusion event was observed in seven of the eight species, with *B. alternifolia* as the notable exception. This divergence may be a significant factor for its speciation from the other species. This finding also indicates that the PPD process plays a significant role in promoting species diversification. Usually, the reduction of chromosome number critically resulted in the abnormal pairing of gametes, ultimately leading to reproductive isolation (Paliulis and Nicklas, 2000; Luo et al., 2018). Additionally, the eight species showed a lower frequency of non-dysploidy alterations, with dysploidy changes being easily identifiable (Figure 3). Interestingly, a higher frequency of EEJ fusion compared to NCF fusion was observed in most species, suggesting that EEJ fusion may have a competitive advantage over

NCF fusion in the process of karyotype evolution. While a similar phenomenon was also reported in previous studies (Wang et al., 2022a), the reliability of this advantage is still an understudied topic. The construction of the LAK and the elucidation of its evolutionary trajectory address a significant gap in our understanding of chromosome karyotype evolution within Lamiales. Furthermore, the discovery revealed that the genomes of the eight karyotype-conserved species possess more complete ancestral chromosomal structures, which suggests their potential as model organisms for future genomic research in Lamiales.

Genomic fractionation and the role of different modes of gene duplications in driving genome evolution

Following polyploidization events, extensive chromosome rearrangements and large-scale gene loss are prevalent due to the dosage balance, particularly in allopolyploids. In this study, following the construction of the post-LAK, we constructed two sets of subgenomes for the eight representative species, respectively. The subgenomes display biased preservation and subgenome dominance, aligning with the lineage-specific hexaploidization seen in *Lupinus* (Xu et al., 2020). This indicated that the L-WGD event may be an allopolyploid event. After observing the fractionation pattern of duplicated genes in these species, we hypothesized that plant species had undergone WGD events that tend to selectively retain these genes within subgenomes in a complementary manner. This suggests that species that underwent WGD events may optimize their genetic repertoire to achieve a more adaptable genetic system in response to changing environments. REs play important roles in driving genome evolution and regulating gene expression (Kubis et al., 1998; Novák et al., 2020; Liao et al., 2023). In this study, we confirmed that the expansion of REs is a key factor influencing genome size variations, which is consistent with some previous studies, and besides polyploidization and gene duplications, repeat expansion was the main factor in amplifying the genome size (Nishihara, 2019; Shao et al., 2019; Novák et al., 2020; de Lima and Ruiz-Ruano, 2022).

Besides, the investigation of different modes of gene duplications across 16 subgenomes revealed that the subgenome 22A exhibited a higher number of duplicate genes than subgenome 22B. This phenomenon shows that gene duplication may play important roles in driving subgenome dominance. Distribution of gene duplication modes across several larger subfamilies in the phylogenetic tree of the *L. philippensis* CYPs superfamily. This diverse distribution also indicates the duplicated gene as a significant force in expanding the gene family (Liu et al., 2023).

Data availability statement

The raw genome sequencing data of *L. philippensis* are available at the National Genomics Data Center (<https://ngdc.cnpc.ac.cn/>) under BioProject number PRJCA010538 (CRA013614). All data are available from the corresponding author upon request.

Author contributions

B-ZC: Writing – original draft, Writing – review & editing, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation. D-WL: Writing – review & editing, Formal analysis. K-YL: Writing – review & editing. S-TJ: Writing – review & editing. XD: Writing – review & editing. W-BW: Writing – review & editing. X-ZL: Writing – review & editing. T-TH: Writing – review & editing. Y-HL: Writing – review & editing. D-ZG: Writing – review & editing. X-TL: Writing – review & editing. S-CD: Writing – review & editing. Y-FZ: Writing – review & editing. WC: Writing – review & editing. YD: Conceptualization, Data curation, Formal analysis, Funding acquisition, Writing – review & editing. W-BY: Conceptualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation of China (31870196, 32371700), the Key Programs of Yunnan Province, China (202101BC070003), the Hainan Province Science and Technology Special Fund (ZDYF2023RDYL01), the Hainan Institute of National Park Fund (KY-24ZK02), and Yunnan Revitalization Talent Support Program “Young Talent” and “Innovation Team” Projects.

References

- Aggarwal, G., and Ramaswamy, R. (2002). Ab initio gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14. doi: 10.1007/bf02703679
- 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
- Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573
- Birchler, J. A., and Han, F. (2018). Barbara McClintock's unsolved chromosomal mysteries: parallels to common rearrangements and karyotype evolution. *Plant Cell* 30, 771–779. doi: 10.1105/tpc.17.00989
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. doi: 10.1093/molbev/msab293
- Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinf. (Oxford England)* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348
- Carta, A., Bedini, G., and Peruzzi, L. (2020). A deep dive into the ancestral chromosome number and genome size of flowering plants. *New Phytol.* 228, 1097–1106. doi: 10.1111/nph.16668
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334
- Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* doi: 10.1002/0471250953.bi0410s05
- Cheng, F., Wu, J., Cai, X., Liang, J., Freeling, M., and Wang, X. (2018). Gene retention, fractionation and subgenome differences in polyploid plants. *Nat. Plants* 4, 258–268. doi: 10.1038/s41477-018-0136-7
- Cheng, T. C., Wu, J., Wu, Y., Chilukuri, R. V., Huang, L., Yamamoto, K., et al. (2017). Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. *Nat. Ecol. Evol.* 1, 1747–1756. doi: 10.1038/s41559-017-0314-4
- Clark, J. W., and Donoghue, P. C. J. (2017). Constraining the timing of whole genome duplication in plant evolutionary history. *Proc. Biol. Sci.* 284, 20170912. doi: 10.1098/rspb.2017.0912
- Clo, J. (2022). Polyploidization: Consequences of genome doubling on the evolutionary potential of populations. *Am. J. Bot.* 109, 1213–1220. doi: 10.1002/ajb2.16029
- Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711
- Cui, L. Y., Wall, P., Leebens-Mack, J., Lindsay, B., Soltis, D., Doyle, J., et al. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Res.* 16, 738–749. doi: 10.1101/gr.4825606
- Damas, J., Kim, J., Farre, M., Griffin, D. K., and Larkin, D. M. (2018). Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol.* 19, 155. doi: 10.1186/s13059-018-1544-8
- Danecek, P., Bonfield, J., Liddle, J., Marshall, J., Ohan, V., Pollard, M., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008. doi: 10.1093/gigascience/giab008
- de Lima, L. G., and Ruiz-Ruano, F. J. (2022). In-depth satelitome analyses of 37 drosophila species illuminate repetitive DNA evolution in the drosophila genus. *Genome Biol. Evol.* 14, evac064. doi: 10.1093/gbe/evac064
- Dodsworth, S., Chase, M. W., and Leitch, A. R. (2015). Is post-polyploidization diploidization the key to the evolutionary success of angiosperms? *Botanical J. Linn. Soc.* 180, 1–5. doi: 10.1111/boj.12357%[Botanical]JournaloftheLinneanSociety
- Dudchenko, O., Batra, S., Omer, A., Nyquist, S., Hoeger, M., Durand, N., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Sci. (New York N.Y.)* 356, 92–95. doi: 10.1126/science.aal3327

Acknowledgments

We are grateful to Mr. Bing Hao and Ms. Shuang Ye for their help in collecting samples. We are grateful to Ms. Yun-bin Pan for his help in data analysis.

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpls.2024.1444234/full#supplementary-material>

- Durand, N. C., Shamim, S., Machol, I., Rao, S., Huntley Miriam, H., Lander, E., et al. (2016). Juicer provides a one-click system for analyzing loop-resolution hi-C experiments. *Cell Syst.* 3, 95–98. doi: 10.1016/j.cels.2016.07.002
- Edger, P. P., Smith, R., McKain, M., Cooley, A., Vallejo-Marin, M., Yuan, Y.-W., et al. (2017). Subgenome dominance in an interspecific hybrid, synthetic allopolyploid, and a 140-year-old naturally established neo-allopolyploid monkeyflower. *Plant Cell* 29, 2150–2167. doi: 10.1105/tpc.17.00010
- Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf.* 9, 18. doi: 10.1186/1471-2105-9-18
- Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi: 10.1186/s13059-019-1832-y
- Fajardo, D., Saint Jean, R., and Lyons, P. J. (2023). Acquisition of new function through gene duplication in the metalloprotease family. *Sci. Rep.* 13, 2512. doi: 10.1038/s41598-023-29800-9
- Feng, C., Wang, J., Wu, L., Kain, H., Yang, L., Feng, C., et al. (2020). The genome of a cave plant, *Primulina huaijiensis*, provides insights into adaptation to limestone karst habitats. *New Phytol.* 227, 1249–1263. doi: 10.1111/nph.16588
- Feng, Y., Wang, Z., Xiao, Q., Teng, J., Wang, J., Yu, Z., et al. (2024). A likely paleo-autotetraploidization event shaped the high conservation of Nyssaceae genome. *Hortic. Plant J.* doi: 10.1016/j.hpj.2022.09.010
- Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367
- Grabherr, M. G., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883
- Guo, S., Zhang, J., Sun, H., Salse, J., Lucas, W. J., Zhang, H., et al. (2013). The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* 45, 51–58. doi: 10.1038/ng.2470
- Haas, B. J., Delcher, A., Mount, S., Wortman, J., Smith, R., Hannick, L., et al. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770
- Haas, B. J., Salzberg, S., Zhu, W., Pertea, M., Allen, J., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7
- Hamilton, J. P., Godden, G., Lanier, T. E., Bhat, W. W., Kinser, T. J., Vaillancourt, B., et al. (2020). Generation of a chromosome-scale genome assembly of the insect-repellent terpenoid-producing Lamiaceae species, *Callicarpa americana*. *GigaScience* 9, g10093. doi: 10.1093/gigascience/giaa093
- Hoang, N. V., Sogbohossou, O., Xiong, W., Simpson, C., Singh, P., Walden, N., et al. (2023). The Gynandropsis gynandra genome provides insights into whole-genome duplications and the evolution of C4 photosynthesis in Cleomaceae. *Plant Cell* 35, 1334–1359. doi: 10.1093/plcell/koad018
- Hu, L., Xu, Z., Wang, M., Fan, R., Yuan, D., Wu, B., et al. (2019). The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. *Nat. Commun.* 10, 4702. doi: 10.1038/s41467-019-12607-6
- Huang, H., Wang, C., Pei, S., and Wang, Y. (2023). A chromosome-level reference genome of an aromatic medicinal plant *Adenosma buchneroides*. *Sci. Data* 10, 660. doi: 10.1038/s41597-023-02571-8
- Jiang, N., Dong, L. N., Yang, J. B., Tan, Y., Wang, H., Randle, C., et al. (2022). Herbarium phylogenomics: Resolving the generic status of the enigmatic Pseudobartsia (Orobanchaceae). *J. Systematics Evol.* 60, 1218–1228. doi: 10.1111/jse.12829
- Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature* 473, 97–100. doi: 10.1038/nature09916
- Julca, I., Marcet-Houben, M., Vargas, P., and Gabaldon, T. (2018). Phylogenomics of the olive tree (*Olea europaea*) reveals the relative contribution of ancient allo- and autopolyploidization events. *BMC Biol.* 16, 15. doi: 10.1186/s12915-018-0482-y
- Jurka, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends genetics: TIG* 16, 418–420. doi: 10.1016/s0168-9525(00)02093-x
- Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285
- Karoonuthaisiri, N., Anghong, P., Uengwetwanit, T., Pootakham, W., Sittikankaw, K., Sonthirod, C., et al. (2020). Optimization of high molecular weight DNA extraction methods in shrimp for a long-read sequencing platform. *PeerJ* 8, e10340. doi: 10.7717/peerj.10340
- Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010
- Kellogg, E. A. (2016). Has the connection between polyploidy and diversification actually been tested? *Curr. Opin. Plant Biol.* 30, 25–32. doi: 10.1016/j.pbi.2016.01.002
- Kong, X., Zhang, Y., Wang, Z., Bao, S., Feng, Y., Wang, J., et al. (2023). Two-step model of paleohexaploidy, ancestral genome reshuffling and plasticity of heat shock response in Asteraceae. *Horticulture Res.* 10, uhad073. doi: 10.1093/hr/uhad073
- Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinf.* 5, 59. doi: 10.1186/1471-2105-5-59
- Kubis, S., Schmidt, T., and Heslop-Harrison, J. S. (1998). Repetitive DNA elements as a major component of plant genomes. *Ann. Bot.* 82, 45–55. doi: 10.1006/anbo.1998.0779
- Landis, J. B., Soltis, D., Li, Z., Marx, H., Barker, M., Tank, D., et al. (2018). Impact of whole-genome duplication events on diversification rates in angiosperms. *Am. J. Bot.* 105, 348–363. doi: 10.1002/ajb2.1060
- Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* 52, W78–W82. doi: 10.1093/nar/gkac268
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinf. (Oxford England)* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, Z., Defoort, J., Tasdighian, S., Maere, S., Van de Peer, Y., De Smet, R., et al. (2016). Gene duplicability of core genes is highly consistent across all angiosperms. *Plant Cell* 28, 326–344. doi: 10.1105/tpc.15.00877
- Li, X., Feng, T., Randle, C., and Schneeweiss, G. (2019). Phylogenetic relationships in orobanchaceae inferred from low-copy nuclear genes: consolidation of major clades and identification of a novel position of the non-photosynthetic orobanche clade sister to all other parasitic orobanchaceae. *Front. Plant Sci.* 10. doi: 10.3389/fpls.2019.00902
- Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., et al. (2023). Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* 6, 954. doi: 10.1038/s42003-023-05322-y
- Liu, X., Gong, Q., Zhao, C., Wang, D., Ye, X., Zheng, G., et al. (2023). Genome-wide analysis of cytochrome P450 genes in Citrus clementina and characterization of a CYP gene encoding flavonoid 3'-hydroxylase. *Horticulture Res.* 10, uhac283. doi: 10.1093/hr/uhac283
- Lovell, J. T., MacQueen, A. H., Mamidi, S., Bonnette, J., Jenkins, J., Napier, J. D., et al. (2021). Genomic mechanisms of climate adaptation in polyploid bioenergy switchgrass. *Nature* 590, 438–43+. doi: 10.1038/s41586-020-03127-1
- Luo, J., Sun, X., Cormack, B. P., and Boeke, J. D. (2018). Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* 560, 392–396. doi: 10.1038/s41586-018-0374-x
- Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinf. (Oxford England)* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315
- Mandáková, T., MacQueen, A. H., Mamidi, S., Bonnette, J., Jenkins, J., and Napier, J. D. (2017). Multispeed genome diploidization and diversification after an ancient allopolyploidization. *Mol. Ecol.* 26, 6445–6462. doi: 10.1111/mec.14379
- Mandáková, T., and Lysak, M. A. (2018). Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42, 55–65. doi: 10.1016/j.pbi.2018.03.001
- Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinf. (Oxford England)* 27, 764–770. doi: 10.1093/bioinformatics/btr011
- Mayrose, I., and Lysak, M. A. (2021). The evolution of chromosome numbers: mechanistic models and experimental approaches. *Genome Biol. Evol.* 13, evaa220.
- Mendes, F. K., Vanderpool, D., Fulton, B., and Hahn, M. W. (2021). CAFE 5 models variation in evolutionary rates among gene families. *Bioinf. (Oxford England)* 36, 5516–5518. doi: 10.1093/bioinformatics/btaa1022
- Morin, S. J., Eccles, J., Iturriaga, A., and Zimmerman, R. S. (2017). Translocations, inversions and other chromosome rearrangements. *Fertility sterility* 107, 19–26.
- Mulcahy, D. G., Noonan, B., Moss, T., Townsend, T., Reeder, T., Sites, J. J., et al. (2012). Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. *Mol. Phylogenet. Evol.* 65, 974–991. doi: 10.1016/j.ympev.2012.08.018
- Murat, F., Armero, A., Pont, C., Klopp, C., and Salse, J. (2017). Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* 49, 490–496. doi: 10.1038/ng.3813
- Mutuku, J. M., Cui, S., Yoshida, S., and Shirasu, K. (2021). Orobanchaceae parasite-host interactions. *New Phytol.* 230, 46–59. doi: 10.1111/nph.17083
- Nishihara, H. (2019). Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes Genet. Syst.* 94, 269–281. doi: 10.1266/ggs.19-00029
- Novák, P., Guignard, M. S., Neumann, P., Kelly, L. J., Mlinarec, J., Kobližková, A., et al. (2020). Repeat-sequence turnover shifts fundamentally in species with large genomes. *Nat. Plants* 6, 1325–1329. doi: 10.1038/s41477-020-00785-x
- Ou, S., and Jiang, N. (2018). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* 176, 1410–1422. doi: 10.1104/pp.17.01310
- Ou, S., and Jiang, N. (2019). LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mobile DNA* 10, 48. doi: 10.1186/s13100-019-0193-0
- Padmarasu, S., Himmelbach, A., Mascher, M., and Stein, N. (2019). *In situ* hi-C for plants: an improved method to detect long-range chromatin interactions. *Methods Mol. Biol. (Clifton N.J.)* 1933, 441–472. doi: 10.1007/978-1-4939-9045-0_28

- Paliulis, L. V., and Nicklas, R. B. (2000). The reduction of chromosome number in meiosis is determined by properties built into the chromosomes. *J. Cell Biol.* 150, 1223–1232. doi: 10.1083/jcb.150.6.1223
- Paterson, A. H., Bowers, J. E., and Chapman, B. A. (2004). Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *PNAS* 101, 9903–9908. doi: 10.1073/pnas.0307901101
- Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., and Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Res.* 46, W200–W204. doi: 10.1093/nar/gky448
- Pryszcz, L. P., and Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res.* 44, e113. doi: 10.1093/nar/gkw294
- Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., et al. (2019). Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* 20, 38. doi: 10.1186/s13059-019-1650-2
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinf. (Oxford England)* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rai, A., Hirakawa, H., Nakabayashi, R., Kikuchi, S., Hayashi, K., Rai, M., et al. (2021). Chromosome-level genome assembly of *Ophiorrhiza pumila* reveals the evolution of camptothecin biosynthesis. *Nat. Commun.* 12, 405. doi: 10.1038/s41467-020-20508-2
- Ranaldo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi: 10.1038/s41467-020-14998-3
- Rastogi, S., and Ohri, D. (2019). Karyotype evolution in cycads. *Nucleus* 63, 131–141. doi: 10.1007/s12327-019-00302-2
- Ren, R., Wang, H., Guo, C., Zhang, N., Zeng, L., Chen, Y., et al. (2018). Widespread whole genome duplications contribute to genome complexity and species diversity in angiosperms. *Mol. Plant* 11, 414–428. doi: 10.1016/j.molp.2018.01.002
- Ren, L. H., Huang, W., and Cannon, T. E. (2019). Reconstruction of ancestral genome reveals chromosome evolution history for selected legume species. *New Phytol.* 223, 2090–2103. doi: 10.1111/nph.15770
- Schubert, I., and Lysak, M. A. (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* 27, 207–216. doi: 10.1016/j.tig.2011.03.004
- Sensalari, C., Maere, S., and Lohaus, R. (2022). ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinf. (Oxford England)* 38, 530–532. doi: 10.1093/bioinformatics/btab602
- Seppely, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol. (Clifton N.J.)* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14
- Shao, F., Han, M., and Peng, Z. (2019). Evolution and diversity of transposable elements in fish genomes. *Sci. Rep.* 9, 15399. doi: 10.1038/s41598-019-51888-1
- Soltis, D. E., Albert, V., Leebens-Mack, J., Bell, C., Paterson, A., Zheng, C., et al. (2009). Polyploidy and angiosperm diversification. *Am. J. Bot.* 96, 336–348. doi: 10.3732/ajb.0800079
- Soltis, P. S., and Soltis, D. E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.* 30, 159–165. doi: 10.1016/j.pbi.2016.03.015
- Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinf. (Oxford England)* 24, 637–644. doi: 10.1093/bioinformatics/btn013
- Stebbins, G. L. Jr. (1947). Types of polyploids; their classification and significance. *Adv. Genet.* 1, 403–429. doi: 10.1016/s0065-2660(08)60490-3
- Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., et al. (2022). WGDf: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol. Plant* 15, 1841–1851. doi: 10.1016/j.molp.2022.10.018
- Sun, Y., Liu, Y., Shi, J., Wang, L., Liang, C., Yang, J., et al. (2023). Biased mutations and gene losses underlying diploidization of the tetraploid broomcorn millet genome. *Plant J.* 113, 787–801. doi: 10.1111/tpj.16085
- Sun, P., Lu, Z., Wang, Z., Wang, S., Zhao, K., Mei, D., et al. (2024). Subgenome-aware analyses reveal the genomic consequences of ancient allopolyploid hybridizations throughout the cotton family. *Proc. Natl. Acad. Sci.* 121, e2313921121. doi: 10.1073/pnas.2313921121
- Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi: 10.1093/nar/gkl315
- Tang, H., Bowers, J., Wang, X., Ming, R., Alam, M., and Paterson, A. (2008). Synteny and collinearity in plant genomes. *Sci. (New York N.Y.)* 320, 486–488. doi: 10.1126/science.1153917
- Tank, D. C., Eastman, J. M., Pennell, M. W., Soltis, P. S., Soltis, D. E., Hinchliff, C. E., et al. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. *New Phytol.* 207, 454–467. doi: 10.1111/nph.13491
- Van de Peer, Y., Mizrahi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet.* 18, 411–424. doi: 10.1038/nrg.2017.26
- Vanneste, K., Baele, G., Maere, S., and Van de Peer, Y. (2014). Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* 24, 1334–1347. doi: 10.1101/gr.168997.113
- Wang, X., Jin, D., Wang, Z., Guo, H., Zhang, L., Wang, L., et al. (2015). Telomeric centromere repatterning determines recurring chromosome number reductions during the evolution of eukaryotes. *New Phytol.* 205, 378–389. doi: 10.1111/nph.12985
- Wang, J. Q., Yuan, M., Feng, Y., Zhang, Y., Bao, S., Hao, Y., et al. (2022a). A common whole-genome paleotetraploidization in Cucurbitales. *Plant Physiol.* 190, 2430–2448. doi: 10.1093/plphys/kiac410
- Wang, L. F., Sun, X., Peng, Y., Chen, K., Wu, S., Guo, Y., et al. (2022b). Genomic insights into the origin, adaptive evolution, and herbicide resistance of *Leptochloa chinensis*, a devastating tetraploid weedy grass in rice fields. *Mol. Plant* 15, 1045–1058. doi: 10.1016/j.molp.2022.05.001
- Wang, Z. Y., Li, Y., Sun, P., Zhu, M., Wang, D., Lu, Z., et al. (2022c). A high-quality *Buxus austro-yunnanensis* (Buxales) genome provides new insights into karyotype evolution in early eudicots. *BMC Biol.* 20, 216. doi: 10.1186/s12915-022-01420-1
- Wang, X. Y., Shi, X. L., Hao, B. L., Ge, S., and Luo, J. C. (2005). Duplication and DNA segmental loss in the rice genome: implications for diploidization. *New Phytol.* 165, 937–946. doi: 10.1111/j.1469-8137.2004.01293.x
- Williams, P. A., Cosme, J., Sridhar, V., Johnson, E. F., and McRee, D. E. (2000). Mammalian microsomal cytochrome P450 monooxygenase: structural adaptations for membrane binding and functional diversity. *Mol. Cell* 5, 121–131. doi: 10.1016/S1097-2765(00)80408-6
- Wu, S. D., Han, B. C., and Jiao, Y. N. (2020). Genetic contribution of paleopolyploidy to adaptive evolution in angiosperms. *Mol. Plant* 13, 59–71. doi: 10.1016/j.molp.2019.10.012
- Xu, W., Zhang, Q., Yuan, W., Xu, F., Muhammad Aslam, M., Miao, R., et al. (2020). The genome evolution and low-phosphorus adaptation in white lupin. *Nat. Commun.* 11, 1069. doi: 10.1038/s41467-020-14891-z
- Xu, Y., Zhang, J., Ma, C., Lei, Y., Shen, G., Jin, J., et al. (2022). Comparative genomics of orobanchaceous species with different parasitic lifestyles reveals the origin and stepwise evolution of plant parasitism. *Mol. Plant* 15, 1384–1399. doi: 10.1016/j.molp.2022.07.007
- Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088
- Zhang, K., Wang, X., and Cheng, F. (2019). Plant polyploidy: origin, evolution, and its influence on crop domestication. *Hortic. Plant J.* 5, 231–239. doi: 10.1016/j.hpj.2019.11.003
- Zhao, Q. Z., Meng, Y., Wang, P., Qin, X., Cheng, C., Zhou, J., et al. (2021). Reconstruction of ancestral karyotype illuminates chromosome evolution in the genus *Cucumis*. *Plant J.* 107, 1243–1259. doi: 10.1111/tpj.15381
- Zhu, Q., Wang, Y., Yao, N., Ni, X., Wang, C., Wang, M., et al. (2023a). Chromosome-level genome assembly of an endangered plant *Prunus mongolica* using PacBio and Hi-C technologies. *DNA Res.* 30, dsad012. doi: 10.1093/dnares/dsad012
- Zhu, S., Zhang, Y., Copsy, L., Han, Q., Zheng, D., Coen, E., et al. (2023b). The snapdragon genomes reveal the evolutionary dynamics of the S-locus supergene. *Mol. Biol. Evol.* 40, msad080. doi: 10.1093/molbev/msad080
- Zhuang, W. J., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., et al. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* 51, 865–86+. doi: 10.1038/s41588-019-0402-2



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Jean-David Rochaix,
University of Geneva, Switzerland
YanJun Jing,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE

Dayong Li

✉ lidayong@nercv.org

Xifeng Chen

✉ xfchen@zjnu.cn

[†]These authors share first authorship

RECEIVED 08 June 2024

ACCEPTED 31 July 2024

PUBLISHED 19 August 2024

CITATION

Zheng M, Wang X, Luo J, Ma B, Li D and
Chen X (2024) The pleiotropic functions of
GOLDEN2-LIKE transcription factors in plants.
Front. Plant Sci. 15:1445875.
doi: 10.3389/fpls.2024.1445875

COPYRIGHT

© 2024 Zheng, Wang, Luo, Ma, Li and Chen.
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.

The pleiotropic functions of GOLDEN2-LIKE transcription factors in plants

Mengyi Zheng^{1†}, Xinyu Wang^{1†}, Jie Luo¹, Bojun Ma¹,
Dayong Li^{2*} and Xifeng Chen^{1*}

¹College of Life Sciences, Zhejiang Normal University, Jinhua, China, ²National Engineering Research Center for Vegetables, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Science, Beijing, China

The regulation of gene expression is crucial for biological plant growth and development, with transcription factors (TFs) serving as key switches in this regulatory mechanism. GOLDEN2-LIKE (GLK) TFs are a class of functionally partially redundant nuclear TFs belonging to the GARP superfamily of MYB TFs that play a key role in regulating genes related to photosynthesis and chloroplast biogenesis. Here, we summarized the current knowledge of the pleiotropic roles of GLKs in plants. In addition to their primary functions of controlling chloroplast biogenesis and function maintenance, GLKs have been proven to regulate the photomorphogenesis of seedlings, metabolite synthesis, flowering time, leaf senescence, and response to biotic and abiotic stress, ultimately contributing to crop yield. This review will provide a comprehensive understanding of the biological functions of GLKs and serve as a reference for future theoretical and applied studies of GLKs.

KEYWORDS

GOLDEN2-LIKEs (GLKs), transcription factor, function, signalling pathway, in plants

Introduction

GOLDEN2-LIKEs (GLKs) are plant-specific transcription factors (TFs) involved in multiple biological processes in plants (Chen et al., 2016; Lambret-Frotte et al., 2023). GLKs are members of the GARP superfamily, containing a nuclear localization signal, a DNA-binding domain (DBD), a proline-rich domain and a GLK/C-terminal (GCT) box (Riechmann et al., 2000; Safi et al., 2017). The DBD consists of three α -helices followed by a highly conserved motif of AREAEAA, which confers specific characteristics to GLKs and distinguishes GLKs from other GARP members (Fitter et al., 2002). To date, GLKs are widespread in land plants, and the last common ancestor of GLKs might be from Embryophyta (Wang et al., 2013; Hernández-Verdeja and Lundgren, 2023). GLKs are demonstrated to be the key regulators for chloroplast biogenesis from lower plants to higher plants (Table 1; Figure 1). Additionally, mounting evidence shows that the GLKs

TABLE 1 Informations and functions of GLKs in plants.

Function	Plant souce	Gene name	Defend against targets	Method	Overexpression host plants	Governance mode	Reference
Chloroplast development	<i>Zea mays</i> (Maize)	<i>ZmGLK1/2</i>	/	OE, KO	Rice	+	(Li et al., 2020b; Yeh et al., 2022)
	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>		OE, KO	Arabidopsis, Tomato	+	(Fitter et al., 2002; Waters et al., 2009; Kobayashi et al., 2012; Powell et al., 2012)
	<i>Physcomitrium patens</i> (Moss)	<i>PpGLK1/2</i>		Homologous recombination	/	+	(Yasumura et al., 2005)
	<i>Oryza sativa</i> (Rice)	<i>OsGLK1/2</i>		OE, KO	Rice	+	(Nakamura et al., 2009; Wang et al., 2013)
	<i>Solanum lycopersicum</i> (Tomato)	<i>SlGLK1/2</i>		OE, KO	Tomato	+	(Nguyen et al., 2014; Niu et al., 2022)
	<i>Capsicum annuum</i> (Pepper)	<i>CaGLK2</i>		Co-localized with <i>pc10</i>	/	+	(Brand et al., 2014)
	<i>Brassica napus</i> (Rapeseed)	<i>BnaGLK1</i>		OE	<i>Brassica napus</i>	+	(Pan et al., 2017; Zhang et al., 2024a)
	<i>Arachis hypogaea</i> (Peanut)	<i>AhGLK1</i>		OE, RNAi	Peanut	+	(Liu et al., 2018, 2020)
	<i>Prunus persica</i> (Peach)	<i>PpGLK1</i>		OE, VIGS	Arabidopsis	+	(Chen et al., 2018)
	<i>Actinidia chinensis</i> (Kiwifruit)	<i>AchGLK</i>		OE	Tomato	+	(Li et al., 2018)
	<i>Malus domestica</i> (Apple)	<i>MpGLK1</i>		OE	Arabidopsis	+	(An et al., 2019; Yang et al., 2023)
	<i>Betula platyphylla</i> × <i>B. pendula</i> (Hybrid birch)	<i>BpGLK1</i>		OE, RNAi	Hybrid birch	+	(Gang et al., 2019)
	<i>Lactuca sativa</i> (Lettuce)	<i>LsGLK</i>		CACTA transposon occurred, Complementation test	/	+	(Zhang et al., 2022b)
	<i>Populus alba</i> × <i>P.berolinensis</i> (Hybrid poplar)	<i>PabGLKs</i>		OE, RNAi	Hybrid poplar		(Li et al., 2021)

(Continued)

TABLE 1 Continued

Function	Plant souce	Gene name	Defend against targets	Method	Overexpression host plants	Governance mode	Reference
	<i>Hordeum vulgare</i> (Barley)	<i>HvGLK1/2</i>		OE, KO	Barley	+	(Taketa et al., 2021)
	<i>Camellia sinensis</i> (Tea plant)	<i>CsGLK1/2</i>		OE	Tomato	+	(Wang et al., 2022)
	<i>Marchantia polymorpha</i> (Liverwort)	<i>MpGLK1</i>		OE, KO	Liverwort	+	(Yelina et al., 2024)
	<i>Raphanus sativus</i> (Radish)	<i>RsGLK2.1</i>		OE, KO	Arabidopsis	+	(Ying et al., 2023)
	<i>Catharanthus roseus</i> (Catharanthus roseus)	<i>CrGLK</i>		VIGS, Chloroplast retrograde signaling inducers	/	+	(Cole-Osborn et al., 2024)
	<i>Liriodendron chinense</i> × <i>L. tulipifera</i> (<i>Liriodendron</i> hybrids)	<i>LhGLK1</i>		OE	Arabidopsis	+	(Qu et al., 2024)
Fruit quality	<i>Solanum lycopersicum</i> (Tomato)	<i>SIGLK1/2</i>		OE	Tomato	+	(Nguyen et al., 2014)
	<i>Oryza sativa</i> (Rice)	<i>OsGLK1/2</i>		OE	Rice	+	(Li et al., 2022c)
	<i>Actinidia chinensis</i> (Kiwifruit)	<i>AchGLK</i>		OE	Tomato	+	(Li et al., 2018)
	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>		OE	Tomato, Arabidopsis		(Powell et al., 2012; Sun et al., 2022)
	<i>Camellia sinensis</i> (Tea plant)	<i>CsGLK1/2</i>		OE	Tomato	+	(Wang et al., 2022)
Flowering	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>		OE, KO	Arabidopsis	–	(Waters et al., 2009; Susila et al., 2023)
	<i>Liriodendron chinense</i> × <i>L. tulipifera</i> (<i>Liriodendron</i> hybrids)	<i>LhGLK1</i>		OE	Arabidopsis	–	(Qu et al., 2024)
Leaf senescence	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>		OE, KO	Arabidopsis	–	(Rauf et al., 2013)
	<i>Brassica napus</i> (Rapeseed)	<i>BnaGLK1a</i>		OE, RNAi	Rapeseed	–	(Zhang et al., 2024a)

(Continued)

TABLE 1 Continued

Function	Plant souce	Gene name	Defend against targets	Method	Overexpression host plants	Governance mode	Reference
Biotic stress responses	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>	<i>Fusarium graminearum</i>	OE	Arabidopsis	+	(Savitch et al., 2007)
			<i>Botrytis cinerea</i>	OE, KO		+	(Murmu et al., 2014)
			<i>Hyaloperonospora arabidopsidis</i> Noco2	OE, KO		+	(Savitch et al., 2007)
			<i>Pseudomonas syringae</i> pv. <i>tomato</i>	KO		–	(Wang et al., 2017a)
			<i>Cucumber mosaic virus</i>	KO		+	(Han et al., 2016)
	<i>Arachis hypogaea</i> (Peanut)	<i>AhGLK1b</i>	<i>Pseudomonas syringae</i> pv. <i>tomato</i>	OE	Peanut	+	(Ali et al., 2020)
	<i>Nicotiana benthamiana</i> (Tobacco)	<i>NbGLK1</i>	<i>Potato virus X</i>	OE	Tobacco	+	(Sukarta et al., 2020)
	<i>Oryza sativa</i> (Rice)	<i>OsGLK1</i>	<i>Rice black-streaked dwarf virus</i>	OE, KO	Rice	+	(Li et al., 2022a)
Abiotic stress responses	<i>Arabidopsis thaliana</i> (Arabidopsis)	<i>AtGLK1/2</i>	Ozone	OE	Arabidopsis	+	(Nagatoshi et al., 2016)
			High light	OE, KO		+	(Zeng et al., 2023; Li et al., 2023b)
			Osmotic and dehydration	OE, KO		–	(Ahmad et al., 2019)
	<i>Arachis hypogaea</i> (Peanut)	<i>AhGLK1</i>	Drought	OE	Arabidopsis	+	(Liu et al., 2018)
	<i>Gossypium hirsutum</i> (Cotton)	<i>GhGLK1</i>	Cold, drought	OE	Arabidopsis	+	(Liu et al., 2021)
	<i>Zea mays</i> (Maize)	<i>ZmGLK1/2</i>	Drought	OE	Rice	+	(Li et al., 2023a)
			High light				(Li et al., 2020b)

OE, Overexpression; RNAi, RNA interference; VIGS, Virus-induced gene silencing; KO, Gene knockout; “+”, Positive regulation; “–”, Negative regulation.

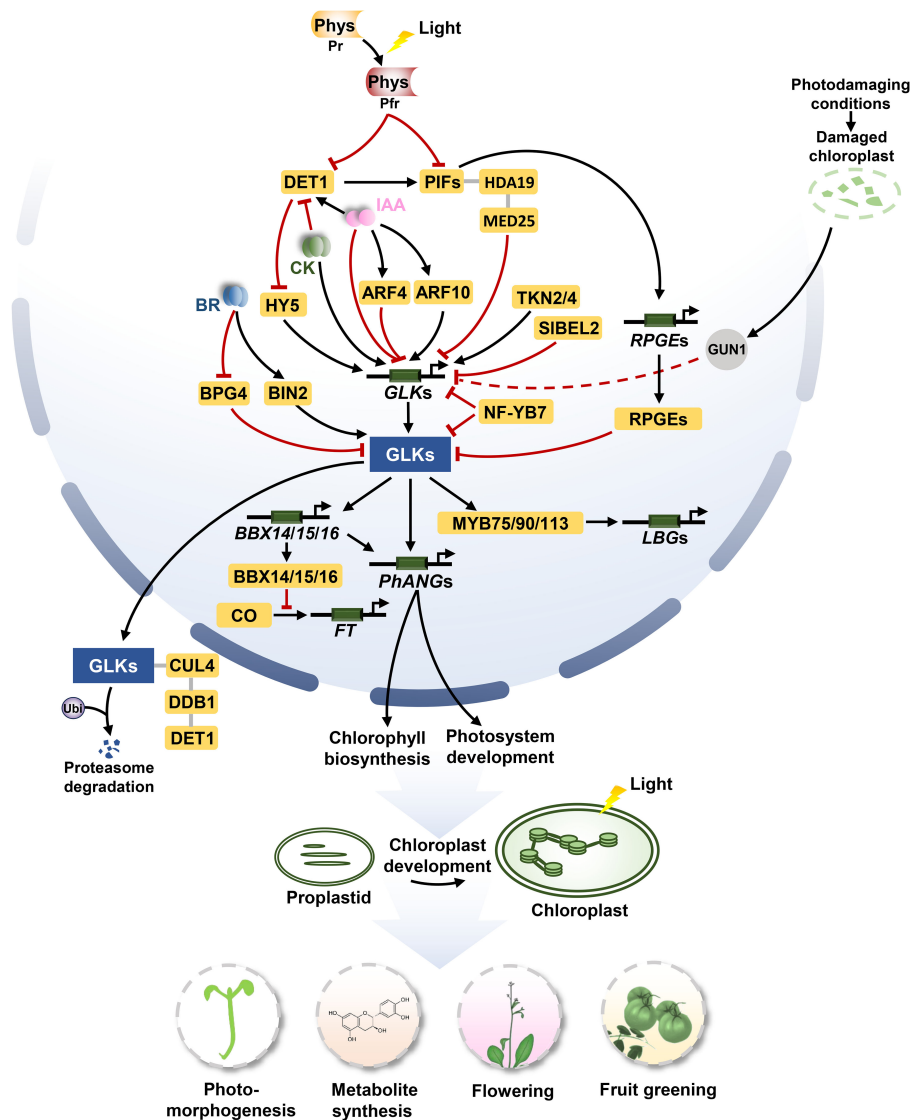


FIGURE 1

The signaling pathways of *GLKs* in regulating chloroplast biogenesis, photomorphogenesis, flowering, and metabolite synthesis. For chloroplast biogenesis, *GLKs* activate the expression of *PhANGs* to promote the development of chloroplast. *TKN2* and *TKN4* activate the expression of *GLK2*, while *BEL2* negatively regulates the expression of *GLK2* to promote the establishment of the 'green shoulder' in tomato fruits. *ARF10* directly induces the expression of *GLK1* and *ARF4* inhibits the transcription of *GLK1*. For photomorphogenesis, activated phytochromes (*Phys*) repress *PIF* and *DET1* under light conditions. *DET1* promotes the stability of *PIF1* proteins, meanwhile, it mediates the proteasome degradation of *GLKs* by interacting with *CUL4* and *DDB1* to form a ubiquitin ligase complex. The *PIF1/PIF3-HDA19-MED25* complex reduces transcriptional repression of *GLK1* under light conditions. Activated *BIN2* phosphorylates and thus stabilizes *GLKs* under light conditions. *BPG4* suppress the transcription activity of *GLKs* via inhibition to their DNA-binding ability. *HY5* binds the promoter of *GLKs*, inducing their activities to promote chloroplast development. Under dark conditions, *PIFs* can directly bind to the *GLK1* promoter to repress the expression of *GLK1*. Moreover, *PIFs* activate the expression of *RPGEs*. *RPGEs* interact with *GLKs* to disrupt the DNA-binding activity of *GLKs*. In photodamaging conditions, the activity of *GUN1* appears to down-regulate the expression of *GLK1* when plastids are dysfunctional. For flowering, *GLKs* directly activate the expression of *BBX14*, *BBX15* and *BBX16*, and the *BBX* proteins physically interact with the circadian clock regulator protein *CO* in the nucleus, which prevents *CO*-mediated *FT* transcription from repressing flowering. For metabolite synthesis, *GLK1* interacts with the MBW complexes *MYB75/90/113* and activates the transcriptional activity to enhance the expression of genes related to anthocyanin-specific biosynthetics including *LBGs*. Arrows and lines with end lines indicate positive regulation and negative regulation, respectively. Grey lines indicate interaction. Dashed arrow represents indirect effects through unknown intermediate factors.

also function in multiple aspects through the entire lifetime of plants, including seedling photomorphogenesis, hormone signalling, leaf senescence, flowering, fruit nutrition and bio- or abiotic stress responses (Table 1; Figures 1, 2). *GLKs* might be a node of signaling networks in plants, which are valuable to research for crop improvement in molecular breeding.

GLKs control chloroplast biogenesis and function maintenance

Chloroplast is an important place for photosynthesis in plants (Jarvis and López-Juez, 2014). Solid evidence indicated that *GLKs* control chloroplast biogenesis by transcriptionally targeting

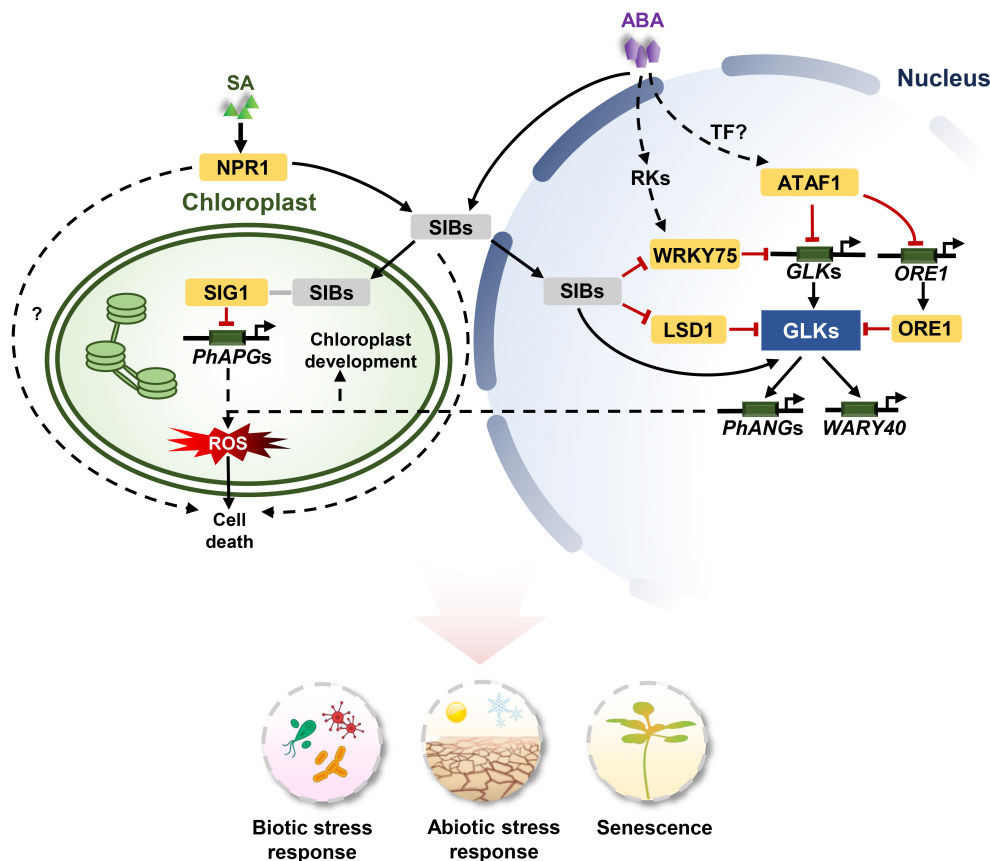


FIGURE 2

The signaling pathways of *GLKs* in stress response and senescence. For biotic stresses, SA-mediated NPR1 activation leads to the expression of *SIB1*. *SIB1* proteins are targeted to both the nucleus and chloroplasts. *SIB1* interacts with *SIG1* to inhibit *PhAPGs* expression in chloroplasts, and *SIB1* activates *GLKs* to induce the expression of *PhANGs* in the nucleus. The uncoupled expression of *PhANGs* and *PhAPGs* leads to an increase of $^1\text{O}_2$ and PQH_2 levels in chloroplasts. The reactive oxygen species (ROS) burst contributes to cell death. The expression of *GLKs* with *SIB1* and functions in cell death. For abiotic stress, *SIBs* are induced by ABA and interact with *WRKY75* to inhibit its transcriptional function. *WRKY75* directly binds to the promoters of *GLKs* to repress their expression. *ATAF1* responds to ABA and suppresses the expression of *GLK1* by directly binding to the promoters of *GLK1* and *ORE1*. *ORE1* interacts with *GLK1* to inhibit its transcriptional activity. *ATAF1* expression is regulated by unknown upstream TFs. ABA activates *GLKs* via core ABA signalling components *PYL/PYRs-PP2Cs-SnRKs*, and subsequently *GLKs* induce the expression of *WRKY40*. Arrows and lines with an end line indicate positive regulation and negative regulation, respectively. Grey lines indicate interaction. Dashed arrows represent indirect effects through unknown intermediate factors.

photosynthesis-related nuclear genes (*PhANGs*), including chlorophyll biosynthesis and photosynthesis-related genes (Waters et al., 2009; Martín et al., 2016). Constitutive expression of *GLKs* could increase chloroplast numbers and chlorophyll content in photosynthetic tissues, such as leaves or fruits (Nguyen et al., 2014), and even in non-photosynthetic tissues such as roots and callus in *Arabidopsis* (*Arabidopsis thaliana*) (Nakamura et al., 2009; Kobayashi et al., 2012). In tomato (*Solanum lycopersicum*), the expression of *GLK2* gradually reduced from the shoulder to the base in fruit, which influences a gradient of chloroplast development of fruit forming the 'green shoulder' fruits (Powell et al., 2012; Nguyen et al., 2014). The TFs KNOTTED1-like Homeobox (KNOX) TKN2 and TKN4 activate the expression of *GLK2* to promote the establishment of 'green shoulder' fruit in tomato (Nadakuduti et al., 2014). However, BEL1-like HOMEODOMAIN 2 (BEL2) affects the formation of 'green shoulder' in tomato fruits by negatively regulating the gradient expression of *GLK2* (Niu et al., 2022). In addition, *GLKs* were affected by AUXIN RESPONSE

FACTORS (ARFs) in regulating chlorophyll accumulation in tomato fruit (Sagar et al., 2013; Yuan et al., 2018). In rice (*Oryza sativa*), a member of the nuclear factor Y (NF-Y) TF family, OsNF-YB7, inactivates the transactivation activity of *GLK1* at multiple regulatory layers to inhibit chlorophyll accumulation in the embryo of rice (Yang et al., 2024). In radish (*Raphanus sativus*), *GLK2* interacts with NUCLEAR FACTOR Y, SUBUNIT A 9a (NF-YA9a) to increase the expression of the chlorophyll biosynthesis gene, *RsHEMA2*, which improves the chloroplast development (Figure 1; Ying et al., 2023).

Interestingly, *GLKs* are functionally redundant in C_3 plants. In *Arabidopsis* and rice, the *glk1* or *glk2* single mutant has no phenotypic difference from the wild type (WT), and the *glk1/glk2* double mutant displayed pale green leaves and abnormal chloroplast structure (Fitter et al., 2002; Wang et al., 2013). However, the functional redundancy of *GLKs* does not exist in the C_4 plant. For instance, maize (*Zea mays*) *glk2* single mutant showed yellow leaves with abnormal chloroplast structure (Rossini

et al., 2001). It is well known that the chloroplasts become different between the C_3 and the C_4 plants, the former has only one type of chloroplast in mesophyll cells (MC), while the latter has two types of chloroplasts in the bundle sheath cells (BSC) and the MC, respectively (Majeran et al., 2009). The development of chloroplasts in the BSC provides an anatomical basis for efficient photosynthesis in C_4 plants (Miyake, 2016). In C_4 plants such as maize and sorghum (*Sorghum bicolor*), *GLK1* expressed much more in MC than that in BSC, while *GLK2* expressed more in BSC contrarily (Wang et al., 2013; John et al., 2014). In addition, the tissue-expression pattern of *GLK1* and *GLK2* are almost similar in Arabidopsis (Supplementary Figure S1), but different in maize (Supplementary Figure S2). It was considered that both *GLK* orthologs retained the ability to induce chloroplast biogenesis and play important roles in regulating the differentiation of chloroplast development in C_4 plants (Rossini et al., 2001), but recent studies showed that *GLK2* adopted a more prominent developmental role, particularly in relation to chloroplast activation in BSC (Lambret-Frotte et al., 2023).

To maintain the functional stability of chloroplasts in plants, the chloroplast-to-nucleus retrograde signalling (RS) is essential for coordinating the expression of *PhANGs* and photosynthesis-associated plastid genes (*PhAPGs*; Pogson et al., 2008). Defective chloroplasts in mutants of plastid protein emphasize coordination between chloroplastic protein processing and nuclear transcription (Chan et al., 2016). GENOMES UNCOUPLED1 (*GUN1*), a chloroplast-localized pentatricopeptide-repeat protein, is a central integrator participating in multiple RS pathways. In photodamaging conditions, the activity of *GUN1* appears to down-regulate the expression of *GLK1* when plastids are dysfunctional (Kakizaki et al., 2010); *GUN1/GLK1* module represses the expression of *B-box structural domain PROTEIN16* (*BBX16*) to regulate the well-established expression of *PhANGs* (Figure 1; Veciana et al., 2022). However, aside from the *GUN1/GLK1* module, studies also showed that the ubiquitin-proteasome system participates in the degradation of Arabidopsis *GLK1* in response to plastid signals in a *GUN1*-independent manner (Tokumaru et al., 2017).

GLKs modulate the photomorphogenesis of seedlings

Seedling photomorphogenesis is coordinately processed as inhibition of hypocotyl elongation, the opening of cotyledon, and chloroplast development when exposed to light. In Arabidopsis, *GLKs* are induced by light (Fitter et al., 2002). The *glk1/glk2* double mutant displayed decreased chlorophyll content, longer hypocotyls and less separated cotyledons (Martín et al., 2016; Alem et al., 2022). PHYTOCHROME-INTERACTING FACTORS (PIFs) are central regulators of photomorphogenesis in plants (Leivar and Monte, 2014). PIFs can form a complex with the histone deacetylase HDA19 and the Mediator subunit MED25, thus attenuating the transcriptional repression of *GLK1* by binding to the PBE motif (CACATG) on *GLK1* promoter in darkness (Martín et al., 2016; Guo et al., 2023), while light-activated phytochrome reverses this activity, thereby inducing *GLKs* expression (Martín et al., 2016).

Interestingly, PIFs can also induce the expression of the *REPRESSOR OF PHOTOSYNTHETIC GENES 1* (*RPGE1*) and *RPGE2* in darkness, and then the RPGEs inhibit the DNA-binding activity of *GLK1* by disrupting its dimerization, revealing another mechanism of PIF-mediated *GLK* repression (Kim et al., 2023). Besides, rice Phytochrome-Interacting Factor-Like1 (*OsPIL1*), a basic helix-loop-helix transcription factor, is also involved in the promotion of chlorophyll biosynthesis (Sakuraba et al., 2017). Moreover, DEETIOLATED 1 (*DET1*), a repressor of light-induced photomorphogenesis, not only promotes the protein stability of *PIF1* (Shi et al., 2015), but also interacts with *GLKs* and promotes the degradation of *GLK* proteins by ubiquitination (Tang et al., 2016; Zhang et al., 2024b). Another regulator of photomorphogenesis, ELONGATED HYPOCOTYL5 (*HY5*) not only directly activates the expression of *GLKs*, but also interacts with the *GLK* proteins, suggesting that *HY5* might first activates the expression of *GLKs* promote chlorophyll biosynthesis and photosystem formation, and then interacts with *GLK* proteins to inhibit hypocotyl elongation (Zhang et al., 2024b). Furthermore, indole-3-acetic acid (IAA) and cytokinin (CK) regulate *GLK2* in the opposing directions at the transcriptional level in a *HY5*-dependent manner to regulate chlorophyll biosynthesis in Arabidopsis roots (Kobayashi et al., 2012).

Additionally, the transcription factor, TEOSINTE BRANCHED 1, CYCLOIDEA, and PROLIFERATING CELL FACTOR 15 (*TCP15*), participates in the expression of *PhANGs* and binds to the same promoter regions of target genes as *GLK1*. It is postulated that *GLK1* helps to recruit *TCP15* for coordinating the expression of cell expansion genes with that of genes involved in the development of the photosynthetic apparatus (Alem et al., 2022). A regulator involved in BR signalling, BRASSINOSTEROID INSENSITIVE2 (*BIN2*), regulates physically interacts with and phosphorylates *GLKs*, and this phosphorylation stabilizes and activates *GLKs* to promote chloroplast development and photomorphogenesis (Zhang et al., 2021). Conversely, BRZINSENSITIVE-PALE GREEN 4 (*BPG4*) inhibits the transcriptional activity of *GLKs* by interacting with the GCT-box of *GLKs* and plays an inhibitory role in regulating chloroplast development and homeostasis (Figure 1; Tachibana et al., 2024).

GLKs participate in the synthesis of metabolites

Photosynthetic products of chloroplasts generally contribute to the accumulation of carbohydrates, lycopene, carotenoids or other nutrient related substances in fruits (Klee and Giovannoni, 2011; Jia et al., 2020). Interestingly, *GLKs* can interact with the G-box Binding Factor (GBF) and activate the transcription of *PHYTOENE SYNTHASE* (*PSY*), promoting the biosynthesis of carotenoids (Sun et al., 2022). Overexpression of the exogenous *GLKs* increases the contents of carbohydrates, carotenoids, and tocopherol (vitamin E) in fruits of tomato (Powell et al., 2012; Nguyen et al., 2014; Lupi et al., 2019). Endosperm-specific overexpression of rice *GLK1* promotes the biosynthesis of carotenoids in the endosperm (Li et al., 2022c). Ectopic overexpression of the *GLK* homolog from

pepper (*Capsicum annuum*), kiwifruit (*Actinidia chinensis*), and tea (*Camellia sinensis*) in tomato resulted in higher levels of carotenoids and sugar in the ripened fruits (Brand et al., 2014; Li et al., 2018; Wang et al., 2022). In addition, GLKs induce the biosynthesis of secondary metabolites including catechin and anthocyanin. *CsGLKs* are also involved in light-regulated catechin accumulation in tea plants by regulating the expression of *CsMYB5b* (Wang et al., 2022). In Arabidopsis, GLK1 interacts with the WD40-BHLH-MYB (MBW) complexes MYB75/90/113 and activates the transcriptional activity to enhance the expression of genes related to anthocyanin-specific biosynthesis including *late biosynthesis genes* (*LBGs*) (Li et al., 2023b). Meanwhile, GLK2 activates the expression of *LBGs* and *TRANSPARENT TESTA GLABRA 1* (*TTG1*) through AtHY5-mediated light signalling and positively regulates anthocyanin biosynthesis in Arabidopsis (Figure 1; Liu et al., 2022; Zeng et al., 2023).

GLKs negatively regulate flowering time and leaf senescence

The flowering time of plants is tightly controlled by endogenous or exogenous signals (Bouché, et al., 2016). It was reported that chloroplasts RS regulated flowering mediated by the floral repressor *FLOWERING LOCUS C* (*FLC*) in Arabidopsis (Feng et al., 2016). GLK1 and GLK2 act as downstream components of the chloroplast RS pathway that negatively regulates flowering time. The *glk1/glk2* double mutant of Arabidopsis displays early flowering, and overexpression of *AtGLK1*, *AtGLK2* or *LhGLK1* in Arabidopsis delayed flowering time (Waters et al., 2009; Qu et al., 2024). GLKs directly activate the expression of *BBX14*, *BBX15* and *BBX16*, and these BBXs proteins physically interact with the circadian clock regulatory *CONSTANS* (*CO*) in the nucleus, which prevent *CO*-mediated *FLOWERING LOCUS T* (*FT*) transcription and repress flowering (Figure 1; Susila et al., 2023).

The chloroplast displays early signs of senescence symptoms, including a decrease in chlorophyll and a decline in photosynthetic efficiency (Soudry et al., 2005). *PIF3*, 4, and 5 are up-regulated during age-triggered and dark-induced leaf senescence, and the accumulation of PIFs protein inhibits the expression of *GLKs* to impair chloroplast development and chlorophyll biosynthesis, leading to leaf senescence (Song et al., 2014). In addition, *GLKs* also respond to abscisic acid (ABA) in regulating plant senescence. The ABA pathway generally promotes leaf senescence, while *GLKs* negatively modulate ABA-mediated leaf senescence. Both *SIBs* and *WRKY75* are upregulated during leaf senescence and induced by ABA. *SIBs* interact with *WRKY75* and thereby repress its transcriptional function, thus negatively regulating ABA-induced leaf senescence in a *WRKY75*-dependent manner. In contrast, *WRKY75* positively modulates ABA-mediated leaf senescence in a *GLK*-dependent manner by directly binding to the W-box (T/CTGACC/T) in the *GLKs* promoter and inhibits their expressions (Zhang et al., 2022a; Lee et al., 2023). In addition, ABA can activate a NAC transcription factor *ATAF1*, which activates *ORESARA1* (*ORE1*) and represses *GLK1* expression by directly binding to the

promoters of both genes. *ORE1* also interacts with *GLKs* to inhibit the transcriptional activity of *GLK1*, resulting in impairing the expression of *GLK* target genes and leaf senescence (Figure 2; Rauf et al., 2013; Garapati et al., 2015). In *Brassica napus*, *GLK1a* has also been shown to directly influence the ABA signalling pathway. Overexpressing *BnGLK1a* delayed the leaf senescence upon ABA treatment (Zhang et al., 2024a).

GLKs are involved in biotic and abiotic stress response

Current studies have shown that *GLKs* participate in the defence response of plants. The *glk1/glk2* double mutant of Arabidopsis showed enhanced resistance to *Pseudomonas syringae* pv. *tomato* and *Hyaloperonospora arabidopsidis* (Wang et al., 2017a). However, overexpression of *AtGLK1* contributes to inducing the expression of *pathogenesis-related* (*PR*) genes, which in turn confers resistance to *Fusarium graminearum* (Savitch et al., 2007). Additionally, overexpression of *AtGLK1* enhances the resistance to *Botrytis cinerea* in a jasmonic acid (*JA*)-independent manner, while increasing the susceptibility to *Hyaloperonospora arabidopsidis* Noco2 in a *JA*-dependant manner (Savitch et al., 2007; Murmu et al., 2014). *GLKs* play positive roles in resistance to cucumber mosaic virus (*CMV*), the Potato virus *X* (*PVX*), the rice black-streaked dwarf virus (*RBSDV*) and the maize rough dwarf disease (*MRDD*) (Han et al., 2016; Sukarta et al., 2020; Li et al., 2022b; Xu et al., 2023). Nevertheless, the virulence protein P69 of Turnip yellow mosaic virus (*TYMV*) interacts with *GLKs* and suppresses *GLKs* transcriptional activity, affecting the normal growth of plants and causing disease symptoms (Ni et al., 2017). Salicylic acid (*SA*) is an important hormone that regulates the defence responses to environmental stresses and against pathogens in plants (Kunkel and Brooks, 2002). *LESION-SIMULATING DISEASE 1* (*LSD1*) is an *SA*-induced cell death regulator and a negative regulator that inhibits the DNA-binding activity of *GLK1* towards its target promoters, and *SIB1* proteins appeared to interrupt the *LSD1*-*GLK* interaction, and the subsequent *SIB1*-*GLK* interaction activated *EX1*-mediated singlet oxygen ($^1\text{O}_2$) signalling, leading to cell death and stress response in plants (Li et al., 2022a).

In addition, *GLKs* actively participate in the response to abiotic stresses. *AhGLK1* upregulates the expression of *AhPORA* during recovery from drought in peanuts (*Arachis hypogaea*), stimulating chlorophyll biosynthesis and photosynthesis to increase the survival rate from drought (Liu et al., 2018). Virus-induced silencing of *GhGLK1* in cotton (*Gossypium hirsutum*) leads to a great impact on growth and yield under drought and cold stress, and *GhGLK1* helps to increase the adaptability of Arabidopsis in drought and cold stress (Liu et al., 2021). Overexpression of maize *GLK* genes in rice improves light harvesting efficiency via Photosystem II (*PSII*), thus buffering the adverse effects of photoinhibition under high or fluctuating light conditions (Li et al., 2023a). In addition, *GLKs* affect ABA sensitivity and ion channel activity of plants to regulate stomatal movements under stresses. The ABA-responsive genes

WRKY40 is regulated by GLKs to increase the sensitivity of seedlings to osmotic stress, and the core ABA signalling components, PYL/PYRs-PP2Cs-SnRKs, possibly act as the intermediary in GLKs-induced *WRKY40* expression (Ahmad et al., 2019). In Arabidopsis, the chimeric repressors for GLKs (GLKs-SRDX) downregulate the genes for inwardly rectifying K⁺ channels and K⁺ channel activity to close the stomata to enhance the tolerance to ozone (Nagatoshi et al., 2016). Recently, the role of GLKs in various abiotic stress responses has been predicted in multiple species through genome-wide analysis, including soybean (*Glycine max*), millet (*Setaria italica*), bamboo (*Phyllostachys edulis*), orange (*Citrus sinensis*) and western balsam poplar (*Populus trichocarpa*) (Alam et al., 2022; Chen et al., 2022; Wu et al., 2022; Xiong et al., 2022; Wu et al., 2023). These facts indicate a broad and conserved function in the abiotic stress response of GLKs in plants, which awaits further validation.

Molecular breeding application of GLKs in crops

Improving plant photosynthesis efficiency is an effective strategy for high-yield breeding in crops. Mounting evidence indicates that manipulation of GLKs achieves yield improvement in plants. In Arabidopsis, leaf-specific and silique wall-specific promoters were used to drive high expression of *AtGLK1*, resulting in enhanced leaf and silique wall photosynthesis and increased seed oil content by 2.88% and 10.75%, respectively (Zhu et al., 2018). In *B. napus*, overexpression of *BnGLK1a* resulted in a 10% increase in the thousand-seed weight of rapeseed (Zhang et al., 2024a). These results suggest that GLKs are promising tools for improving seed yield and oil production in oilseed crops.

Since the photosynthesis efficiency of C₄ plants is much higher than that of C₃ plants (von Caemmerer et al., 2012), the ectopic expression of maize (C₄ plant) *ZmGLKs* was carried in rice (C₃ plant) to improve its yield. The engineering rice plants induced chloroplast development in BSC accompanied by the accumulation of photosynthetic enzymes and intercellular connections (Wang et al., 2017b; Yeh et al., 2022). Overexpression of the *ZmGLK1* and *ZmGLK2* in rice increased the yield by 30% to 40% (Li et al., 2020b), while expression of *ZmGLKs* driven by its native promoter in rice increased the yield by 47% to 70% (Yeh et al., 2022).

Discussion

GLK is a key regulator of chloroplast development. Knockout of GLKs lead to abnormal chloroplast structure but not complete distortion of chloroplast biogenesis (Fitter et al., 2002; Wang et al., 2013), suggesting the existence of other genes which can partly compensate for GLKs function in chloroplast development. Besides, though GLKs are considered to play important roles in regulating the differentiation of chloroplast development in C₄ plants (Rossini et al., 2001), the molecular mechanism remains unclear. Recently, it

was shown that the pleiotropic role of GLKs beyond chloroplast regulation, including photomorphogenesis, synthesis of secondary metabolites, flowering, senescence and response to biotic and abiotic stresses (Table 1). Regarding GLKs being functionally redundant in chloroplast development in C₃ plants, it's natural to think whether GLKs are also redundant in regulating other aspects of life. Clarifying these questions would be helpful in understanding the bio-function of GLK in plants.

As core regulators in plant, GLKs are involved in multiple molecular modes of action including response to upstream genes, binding to downstream target genes and protein-protein interactions. However, so far, some studies only proved the interaction relationship between GLK and target proteins. The specific binding elements still await further research. The expression of GLK can be regulated by the upstream regulators by binding to specific *cis*-elements in the promoter, such as T/CTGACC/T (W-box), CACGTG (G-box) or CACATG (E-box) (Zhang et al., 2022a; Sakuraba et al., 2017). Besides, GLK can also bind to the promotor of target genes downstream to regulate their expression. The highly conserved motif CCAATC is considered a widely shared *cis*-acting element for downstream targets of GLKs (Waters et al., 2009). Comparative cross-species analyses of GLKs have shown that most of the binding sites of GLKs were species-specific (Tu et al., 2022), providing support for further exploration of binding sites rich in downstream targets of GLKs in the future. Furthermore, the DNA-binding domain and GCT-box of GLK proteins are specific binding domains for most regulatory factors. Interestingly, a few proteins also bind to proline-rich regions of GLK proteins, such as LSD1 (Li et al., 2022a). As for the degradation, SIGLK2 is proven a substrate of the CULLIN4 (CUL4) - UV-DAMAGED DNA BINDING PROTEIN 1 (DDB1) - DET1 ubiquitin ligase complex for the proteasome degradation (Tang et al., 2016). However, the ubiquitin-proteasome system is also shown to participate in the degradation of Arabidopsis GLK1 in response to plastid signals (Tokumaru et al., 2017). Would it also be a part of the 'CUL4-DDB1-DET1 degradation pathway'? Further research is needed to clarify their relationship.

In addition, GLKs have shown a rosy application prospect. By regulating the gene expression of GLKs, not only can the photosynthetic efficiency of crops be increased which in turn improves crop yields, but leaf morphogenesis can also be changed. It makes GLKs potentially applicable to agronomic trait improvement, horticultural plant breeding and ornamental plant improvement. However, overexpression of GLKs has certain negative effects. For example, transgenic rice of *ZmG1* driven by the constitutive promoter resulted in reduced seed size and no increase in yield (Yeh et al., 2022). Overexpression of *OsGLK1* in rice causes abnormal tapetum development and low seed setting rates, and also increased endosperm chalkiness of rice grains (Zheng et al., 2022; Li et al., 2022c). To mitigate the potential negative effects, the expression level of GLKs may be tightly regulated by selecting appropriate promoters, or 'Knock-up' by gene-editing techniques (Lu et al., 2021; Wang et al., 2024). Accurate regulation of the expression of GLKs will help improve crop overall quality and bring breakthroughs in agricultural production.

Author contributions

MZ: Writing – original draft, Writing – review & editing. XW: Writing – original draft, Writing – review & editing. JL: Writing – original draft, Writing – review & editing. BM: Writing – review & editing. DL: Project administration, Supervision, Writing – review & editing. XC: Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the National Natural Science Foundation of China (32272096), Innovation and Development Program of the Beijing Vegetable Research Center (KYCX202304) and the Natural Science Foundation of Zhejiang Province (LZ23C130004 and Z24C130016).

Acknowledgments

We thank Dr. Tianhua He (Murdoch University, Australia) for critical reading of this manuscript.

References

- Ahmad, R., Liu, Y., Wang, T. J., Meng, Q., Yin, H., Wang, X., et al. (2019). GOLDEN2-LIKE transcription factors regulate *WRKY40* expression in response to abscisic acid. *Plant Physiol.* 179, 1844–1860. doi: 10.1104/pp.18.01466
- Alam, I., Manghwar, H., Zhang, H., Yu, Q., and Ge, L. (2022). Identification of GOLDEN2-like transcription factor genes in soybeans and their role in regulating plant development and metal ion stresses. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1052659
- Alem, A. L., Ariel, F. D., Cho, Y., Hong, J. C., Gonzalez, D. H., and Viola, I. L. (2022). TCP15 interacts with GOLDEN2-LIKE 1 to control cotyledon opening in *Arabidopsis*. *Plant J.* 110, 748–763. doi: 10.1111/tj.15701
- Ali, N., Chen, H., Zhang, C., Khan, S. A., Gandeka, M., Xie, D., et al. (2020). Ectopic expression of *AhGLK1b* (GOLDEN2-like transcription factor) in *Arabidopsis* confers dual resistance to fungal and bacterial pathogens. *Genes* 11, 343. doi: 10.3390/genes11030343
- An, X. H., Tian, Y., Chen, Y. H., Li, E. M., Li, M., and Cheng, C. G. (2019). Functional identification of apple *MdGLK1* which regulates chlorophyll biosynthesis in *Arabidopsis*. *J. Plant Growth Regul.* 38, 778–787. doi: 10.1007/s00344-018-9889-5
- Bouché, F., Lobet, G., Tocquin, P., and Perilleux, C. (2016). FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* 44, 1167–1171. doi: 10.1093/nar/gkv1054
- Brand, A., Borovsky, Y., Hill, T., Rahman, K. A., Bellalou, A., Van Deynze, A., et al. (2014). *CaGLK2* regulates natural variation of chlorophyll content and fruit color in pepper fruit. *Theor. Appl. Genet.* 127, 2139–2148. doi: 10.1007/s00122-014-2367-y
- Chan, K. X., Phua, S. Y., Crisp, P., McQuinn, R., and Pogson, B. J. (2016). Learning the languages of the chloroplast: retrograde signaling and beyond. *Annu. Rev. Plant Biol.* 67, 25–53. doi: 10.1146/annurev-arplant-043015-111854
- Chen, M., Ji, M., Wen, B., Liu, L., Li, S., Chen, X., et al. (2016). GOLDEN 2-LIKE transcription factors of plants. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01509
- Chen, M., Liu, X., Jiang, S., Wen, B., Yang, C., Xiao, W., et al. (2018). Transcriptomic and functional analyses reveal that *PpGLK1* regulates chloroplast development in peach (*Prunus persica*). *Front. Plant Sci.* 9. doi: 10.3389/fpls.2018.00034
- Chen, H., Qin, L., Tian, J., and Wang, X. (2022). Identification and evolutionary analysis of the *GOLDEN2-LIKE* gene family in Foxtail Millet. *Trop. Plant Biol.* 15, 301–318. doi: 10.1007/s12042-022-09324-8
- Cole-Osborn, L. F., McCallan, S. A., Prifti, O., Abu, R., Sjoelund, V., and Lee-Parsons, C. W. T. (2024). The role of the Golden2-like (GLK) transcription factor in regulating

Conflict of interest

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

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

Publisher's note

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.2024.1445875/full#supplementary-material>

terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *Plant Cell Rep.* 43, 141. doi: 10.1007/s00299-024-03208-9

Feng, P., Guo, H., Chi, W., Chai, X., Sun, X. W., Xu, X. M., et al. (2016). Chloroplast retrograde signal regulates flowering. *PANS* 113, 10708–10713. doi: 10.1073/pnas.1521599113

Fitter, D. W., Martin, D. J., Copley, M. J., Scotland, R. W., and Langdale, J. A. (2002). *GLK* gene pairs regulate chloroplast development in diverse plant species. *Plant J.* 31, 713–727. doi: 10.1046/j.1365-3113x.2002.01390.x

Gang, H., Li, R., Zhao, Y., Liu, G., Chen, S., and Jiang, J. (2019). Loss of *GLK1* transcription factor function reveals new insights in chlorophyll biosynthesis and chloroplast development. *J. Exp. Bot.* 70, 3125–3138. doi: 10.1093/jxb/erz128

Garapati, P., Xue, G. P., Munné-Bosch, S., and Balazadeh, S. (2015). Transcription factor *ATAF1* in *Arabidopsis* promotes senescence by direct regulation of key chloroplast maintenance and senescence transcriptional cascades. *Plant Physiol.* 168, 1122–1139. doi: 10.1104/pp.15.00567

Guo, Q., Jing, Y. J., Gao, Y., Liu, Y. T., Fang, X. F., and Lin, R. C. (2023). The PIF1/PIF3-MED25-HDA19 transcriptional repression complex regulates phytochrome signaling in *Arabidopsis*. *New Phytol.* 240, 1097–1115. doi: 10.1111/nph.19205

Han, X. Y., Li, P. X., Zou, L. J., Tan, W. R., Zheng, T., Zhang, D. W., et al. (2016). *GOLDEN2-LIKE* transcription factors coordinate the tolerance to Cucumber mosaic virus in *Arabidopsis*. *Biochem. Biophys. Res. Commun.* 477, 626–632. doi: 10.1016/j.bbrc.2016.06.110

Hernández-Verdeja, T., and Lundgren, M. R. (2023). GOLDEN2-LIKE transcription factors: A golden ticket to improve crops? *Plants People Planet.* 6, 79–93. doi: 10.1002/ppp3.10412

Jarvis, P., and López-Juez, E. (2014). Biogenesis and homeostasis of chloroplasts and other plastids. *Nat. Rev. Mol. Cell Biol.* 14, 787–802. doi: 10.1038/nrm3702

Jia, T., Cheng, Y., Khan, I., Zhao, X., Gu, T., and Hu, X. (2020). Progress on understanding transcriptional regulation of chloroplast development in fleshy fruit. *Int. J. Mol. Sci.* 21, 6951. doi: 10.3390/ijms21186951

John, C. R., Smith-Unna, R. D., Woodfield, H., Covshoff, S., and Hibberd, J. M. (2014). Evolutionary convergence of cell-specific gene expression in independent lineages of *C₄* grasses. *Plant Physiol.* 165, 62–75. doi: 10.1104/pp.114.238667

Kakizaki, T., Matsumura, H., Nakayama, K., Che, F. S., Terauchi, R., and Inaba, T. (2010). Coordination of plastid protein import and nuclear gene expression by plastid-to-nucleus retrograde signaling. *Plant Physiol.* 151, 1339–1353. doi: 10.1104/pp.109.145987

- Kim, N., Jeong, J., Kim, J., Oh, J., and Choi, G. (2023). Shade represses photosynthetic genes by disrupting the DNA binding of GOLDEN2-LIKE1. *Plant Physiol.* 191, 2334–2352. doi: 10.1093/plphys/kiad029
- Klee, H. J., and Giovannoni, J. J. (2011). Genetics and control of tomato fruit ripening and quality attributes. *Annu. Rev. Genet.* 45, 41–59. doi: 10.1146/annurev-genet-110410-132507
- Kobayashi, K., Baba, S., Obayashi, T., Sato, M., Toyooka, K., Keränen, M., et al. (2012). Regulation of root greening by light and auxin/cytokinin signaling in *Arabidopsis*. *Plant Cell* 24, 1081–1095. doi: 10.1105/tpc.111.092254
- Kunkel, B. N., and Brooks, D. M. (2002). Cross talk between signaling pathways in pathogen defense. *Curr. Opin. Plant Biol.* 5, 325–331. doi: 10.1016/s1369-5266(02)00275-3
- Lambret-Frotte, J., Smith, G., and Langdale, J. A. (2023). *GOLDEN2-like 1* is sufficient but not necessary for chloroplast biogenesis in mesophyll cells of *C₄* grasses. *Plant J.* 117, 416–431. doi: 10.1007/s00299-024-03208-9
- Lee, K. P., Li, M., Li, M., Liu, K., Medina-Puche, L., Qi, S., et al. (2023). Hierarchical regulatory module GENOMES UNCOUPLED1-GOLDEN2-LIKE1/2-WRKY18/40 modulates salicylic acid signaling. *Plant Physiol.* 192, 3120–3133. doi: 10.1093/plphys/kiad251
- Leivar, P., and Monte, E. (2014). PIFs: systems integrators in plant development. *Plant Cell* 26, 56–78. doi: 10.1105/tpc.113.120857
- Li, G., Chen, D., Tang, X., and Liu, Y. (2018). Heterologous expression of kiwifruit (*Actinidia chinensis*) GOLDEN2-LIKE homolog elevates chloroplast level and nutritional quality in tomato (*Solanum lycopersicum*). *Planta* 247, 1351–1362. doi: 10.1007/s00425-018-2853-6
- Li, Z., Gao, J., Wang, B., Xu, J., Fu, X., Han, H., et al. (2022c). Rice carotenoid biofortification and yield improvement conferred by endosperm-specific overexpression of *OsGLK1*. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.951605
- Li, Y., Gu, C., Gang, H., Zheng, Y., Liu, G., and Jiang, J. (2021). Generation of a golden leaf triploid poplar by repressing the expression of *GLK* genes. *Forest. Res.* 1. doi: 10.3389/fpls.2022.952877
- Li, M., Lee, K. P., Liu, T., Vivek, D., Duan, J., Li, M., et al. (2022a). Antagonistic modules, SIB1 and LSD1, regulate photosynthesis-associated nuclear genes via GOLDEN2-LIKE transcription factors in *Arabidopsis*. *Plant Physiol.* 188, 2308–2324. doi: 10.1093/plphys/kiab600
- Li, Y., Lei, W., Zhou, Z., Li, Y., Zhang, D., and Lin, H. (2023b). Transcription factor *GLK1* promotes anthocyanin biosynthesis via an MBW complex-dependent pathway in *Arabidopsis thaliana*. *J. Integr. Plant Biol.* 65, 1521–1535. doi: 10.1111/jipb.13471
- Li, X., Li, J., Wei, S., Gao, Y., Pei, H., Geng, R., et al. (2023a). Maize GOLDEN2-LIKE proteins enhance drought tolerance in rice by promoting stomatal closure. *Plant Physiol.* 194, 774–786. doi: 10.1093/plphys/kiad561
- Li, X., Lin, F., Li, C., Du, L., Liu, Z., Shi, W., et al. (2022b). Golden 2-like transcription factor contributes to the major QTL against rice black-streaked dwarf virus disease. *Theor. Appl. Genet.* 135, 4233–4243. doi: 10.1007/s00122-022-04214-9
- Li, X., Wang, P., Li, J., Wei, S., Yan, Y., Yang, J., et al. (2020b). Maize GOLDEN2-LIKE genes enhance biomass and grain yields in rice by improving photosynthesis and reducing photoinhibition. *Commun. Biol.* 3, 151. doi: 10.1038/s42003-020-0887-3
- Liu, X., Li, L., Li, M., Su, L., Lian, S., Zhang, B., et al. (2018). *AhGLK1* affects chlorophyll biosynthesis and photosynthesis in peanut leaves during recovery from drought. *Sci. Rep.* 8, 2250. doi: 10.1038/s41598-018-20542-7
- Liu, X., Li, L., Zhang, B., Zeng, L., and Li, L. (2020). *AhHDA1*-mediated *AhGLK1* promoted chlorophyll synthesis and photosynthesis regulates recovery growth of peanut leaves after water stress. *Plant Sci.* 294, 110461. doi: 10.1016/j.plantsci.2020.110461
- Liu, J., Mehari, T. G., Xu, Y., Umer, M. J., Hou, Y., Wang, Y., et al. (2021). *GhGLK1* a key candidate gene from GARP family enhances cold and drought stress tolerance in cotton. *Front. Plant Sci.* 12, 759312. doi: 10.3389/fpls.2021.759312
- Liu, D., Zhao, D., Li, X., and Zeng, Y. (2022). *AtGLK2*, an *Arabidopsis* GOLDEN2-LIKE transcription factor, positively regulates anthocyanin biosynthesis via *AtHY5*-mediated light signaling. *Plant Growth Regul.* 96, 79–90. doi: 10.1007/s10725-021-00759-9
- Lu, Y., Wang, J. Y., Chen, B., Mo, S. D., Lian, L., Luo, Y. M., et al. (2021). A donor-DNA-free CRISPR/Cas-based approach to gene knock-up in rice. *Nat. Plants* 7, 1445–1452. doi: 10.1038/s41477-021-01019-4
- Lupi, A. C. D., Lira, B. S., Gramegna, G., Trench, B., Alves, F. R. R., Demarco, D., et al. (2019). *Solanum lycopersicum* GOLDEN 2-LIKE 2 transcription factor affects fruit quality in a light- and auxin-dependent manner. *PLoS One* 14, 212224. doi: 10.1371/journal.pone.0212224
- Majeran, W., and Van Wijk, K. J. (2009). Cell-type-specific differentiation of chloroplasts in *C₄* plants. *Trends Plant Sci.* 14, 100–109. doi: 10.1016/j.tplants.2008.11.006
- Martin, G., Leivar, P., Ludevid, D., Tepperman, J. M., Quail, P. H., and Monte, E. (2016). Phytochrome and retrograde signalling pathways converge to antagonistically regulate a light-induced transcriptional network. *Nat. Commun.* 7, 11431. doi: 10.1038/ncomms11431
- Miyake, H. (2016). Starch accumulation in the bundle sheaths of *C₃* plants: A possible pre-condition for *C₄* photosynthesis. *Plant Cell Physiol.* 57, 890–896. doi: 10.1093/pcp/pcw046
- Murmu, J., Wilton, M., Allard, G., Pandeya, R., Desveaux, D., Singh, J., et al. (2014). *Arabidopsis* GOLDEN2-LIKE (GLK) transcription factors activate jasmonic acid (JA)-dependent disease susceptibility to the biotrophic pathogen *Hyaloperonospora arabidopsidis*, as well as JA-independent plant immunity against the necrotrophic pathogen *Botrytis cinerea*. *Mol. Plant Pathol.* 15, 174–184. doi: 10.1111/mpp.12077
- Nadakuduti, S. S., Holdsworth, W. L., Klein, C. L., and Barry, C. S. (2014). *KNOX* genes influence a gradient of fruit chloroplast development through regulation of *GOLDEN2-LIKE* expression in tomato. *Plant J.* 78, 1022–1033. doi: 10.1111/tpj.12529
- Nagatoshi, Y., Mitsuda, N., Hayashi, M., Inoue, S., Okuma, E., Kubo, A., et al. (2016). GOLDEN2-LIKE transcription factors for chloroplast development affect ozone tolerance through the regulation of stomatal movement. *Proc. Natl. Acad. Sci. U.S.A.* 113, 4218–4223. doi: 10.1073/pnas.1513093113
- Nakamura, H., Muramatsu, M., Hakata, M., Ueno, O., Nagamura, Y., Hirochika, H., et al. (2009). Ectopic overexpression of the transcription factor *OsGLK1* induces chloroplast development in non-green rice cells. *Plant Cell Physiol.* 50, 1933–1949. doi: 10.1093/pcp/pcp138
- Nguyen, C. V., Vrebalov, J. T., Gapper, N. E., Zheng, Y., Zhong, S., Fei, Z., et al. (2014). Tomato *GOLDEN2-LIKE* transcription factors reveal molecular gradients that function during fruit development and ripening. *Plant Cell* 26, 585–601. doi: 10.1105/tpc.113.118794
- Ni, F., Wu, L., Wang, Q., Hong, J., Qi, Y., and Zhou, X. (2017). Turnip Yellow Mosaic Virus P69 interacts with and suppresses GLK transcription factors to cause Pale-Green symptoms in *Arabidopsis*. *Mol. Plant* 10, 764–766. doi: 10.1016/j.molp.2016.12.003
- Niu, X. L., Li, H. L., Li, R., Liu, G. S., Peng, Z. Z., Jia, W., et al. (2022). Transcription factor SIBEL2 interferes with GOLDEN2-LIKE and influences green shoulder formation in tomato fruits. *Plant J.* 112, 982–997. doi: 10.1111/tpj.15989
- Pan, Y. L., Pan, Y., Qu, C. M., Su, C. G., Li, J. H., and Zhang, X. G. (2017). Identification and cloning of *GOLDEN2-LIKE1 (GLK1)*, a transcription factor associated with chloroplast development in *Brassica napus* L. *Genet. Mol. Res.* 16. doi: 10.4238/gmr16018942
- Pogson, B. J., Woo, N. S., Forster, B., and Small, I. D. (2008). Plastid signalling to the nucleus and beyond. *Trends Plant Sci.* 13, 602–609. doi: 10.1016/j.tplants.2008.08.008
- Powell, A. L., Nguyen, C. V., Hill, T., Cheng, K. L., Figueroa-Balderas, R., Aktas, H., et al. (2012). *Uniform ripening* encodes a *Golden2-like* transcription factor regulating tomato fruit chloroplast development. *Science* 336, 1711–1715. doi: 10.1126/science.1222218
- Qu, H. X., Liang, S., Hu, L. F., Yu, L., Liang, P. X., Hao, Z. D., et al. (2024). Overexpression of *Liriodendron Hybrid LhGLK1* in *Arabidopsis* leads to excessive chlorophyll synthesis and improved growth. *Int. J. Mol. Sci.* 25, 6869. doi: 10.3390/ijms25136968
- Rauf, M., Arif, M., Dortay, H., Matallana-Ramirez, L. P., Waters, M. T., Gil Nam, H., et al. (2013). *ORE1* balances leaf senescence against maintenance by antagonizing G2-like-mediated transcription. *EMBO Rep.* 14, 382–388. doi: 10.1038/embor.2013.24
- Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., et al. (2000). *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290, 2105–2110. doi: 10.1126/science.290.5499.2105
- Rossini, L., Cribb, L., Martin, D. J., and Langdale, J. A. (2001). The maize *golden 2* gene defines a novel class of transcriptional regulators in plants. *Plant Cell* 13, 1231–1244. doi: 10.1105/tpc.13.5.1231
- Safi, A., Medici, A., Szponarski, W., Ruffel, S., Lacombe, B., and Krouk, G. (2017). The world according to GARP transcription factors. *Curr. Opin. Plant Biol.* 39, 159–167. doi: 10.1016/j.cpb.2017.07.006
- Sagar, M., Chervin, C., Mila, I., Hao, Y., Roustan, J. P., Benichou, M., et al. (2013). *SIARF4*, an auxin response factor involved in the control of sugar metabolism during tomato fruit development. *Plant Physiol.* 161, 1362–1374. doi: 10.1104/pp.113.213843
- Sakuraba, Y., Kim, E. Y., Han, S. H., Piao, W., An, G., Todaka, D., et al. (2017). Rice Phytochrome-Interacting Factor-Like1 (OSPI1) is involved in the promotion of chlorophyll biosynthesis through feed-forward regulatory loops. *J. Exp. Bot.* 68, 4103–4114. doi: 10.1093/jxb/erx231
- Savitch, L. V., Subramaniam, R., Allard, G. C., and Singh, J. (2007). The *GLK1* ‘regulon’ encodes disease defense related proteins and confers resistance to *Fusarium graminearum* in *Arabidopsis*. *Biochem. Biophys. Res. Commun.* 359, 234–238. doi: 10.1016/j.bbrc.2007.05.084
- Shi, H., Wang, X., Mo, X., Tang, C., Zhong, S., and Deng, X. W. (2015). *Arabidopsis* DET1 degrades HFR1 but stabilizes PIF1 to precisely regulate seed germination. *Proc. Natl. Acad. Sci. U S A* 112, 3817–3822. doi: 10.1073/pnas.1502405112
- Song, Y., Yang, C., Gao, S., Zhang, W., Li, L., and Kuai, B. (2014). Age-triggered and dark-induced leaf senescence require the bHLH transcription factors PIF3, 4, and 5. *Mol. Plant* 7, 1776–1787. doi: 10.1093/mp/ssu109
- Soudry, E., Ulitzur, S., and Gepstein, S. (2005). Accumulation and remobilization of amino acids during senescence of detached and attached leaves: in planta analysis of tryptophan levels by recombinant luminescent bacteria. *J. Exp. Bot.* 56, 695–702. doi: 10.1093/jxb/eri054
- Sukarta, O. C. A., Townsend, P. D., Llewellyn, A., Dixon, C. H., Sloatweg, E. J., Pålsson, L. O., et al. (2020). A DNA-Binding Bromodomain-containing protein interacts with and reduces Rx1-mediated immune response to potato virus X. *Plant Commun.* 1, 100086. doi: 10.1016/j.xplc.2020.100086
- Sun, T., Zeng, S., Wang, X., Owens, L. A., Fe, Z., Zhao, Y., et al. (2022). GLKs directly regulate carotenoid biosynthesis via interacting with GBFs in nuclear condensates in plants. *bioRxiv*. doi: 10.1101/2022.09.09.507346

- Susila, H., Nasim, Z., Gawarecka, K., Jung, J. Y., Jin, S., Youn, G., et al. (2023). Chloroplasts prevent precocious flowering through a *GOLDEN2-LIKE-B-BOX DOMAIN PROTEIN* module. *Plant Commun.* 4, 100515. doi: 10.1016/j.xplc.2023.100515
- Tachibana, R., Abe, S., Marugami, M., Yamagami, A., Akema, R., Ohashi, T., et al. (2024). BPG4 regulates chloroplast development and homeostasis by suppressing GLK transcription factors and involving light and brassinosteroid signaling. *Nat. Commun.* 15, 370. doi: 10.1038/s41467-023-44492-5
- Taketa, S., Hattori, M., Takami, T., Himi, E., and Sakamoto, W. (2021). Mutations in a *Golden2-Like* Gene Cause Reduced Seed Weight in Barley *albino lemma 1* Mutants. *Plant Cell Physiol.* 62, 447–457. doi: 10.1093/pcp/pcab001
- Tang, X., Miao, M., Niu, X., Zhang, D., Cao, X., Jin, X., et al. (2016). Ubiquitin-conjugated degradation of golden 2-like transcription factor is mediated by CUL4-DDB1-based E3 ligase complex in tomato. *New Phytol.* 209, 1028–1039. doi: 10.1111/nph.13635
- Tokumaru, M., Adachi, F., Toda, M., Ito-Inaba, Y., Yazu, F., Hirosawa, Y., et al. (2017). Ubiquitin-proteasome dependent regulation of the golden 2-like 1 transcription factor in response to plastid signals. *Plant Physiol.* 173, 524–535. doi: 10.1104/pp.16.01546
- Tu, X., Ren, S., Shen, W., Li, J., Li, Y., Li, C., et al. (2022). Limited conservation in cross-species comparison of GLK transcription factor binding suggested wide-spread cisrome divergence. *Nat. Commun.* 13, 7632. doi: 10.1038/s41467-022-35438-4
- Veciana, N., Martín, G., Leivar, P., and Monte, E. (2022). BBX16 mediates the repression of seedling photomorphogenesis downstream of the GUN1/GLK1 module during retrograde signalling. *New Phytol.* 234, 93–106. doi: 10.1111/nph.17975
- von Caemmerer, S., Quick, W. P., and Furbank, R. T. (2012). The development of *C₄* rice: current progress and future challenges. *Science* 336, 1671–1672. doi: 10.1126/science.1220177
- Wang, P., Fouracre, J., Kelly, S., Karki, S., Gowik, U., Aubry, S., et al. (2013). Evolution of *GOLDEN2-LIKE* gene function in *C₃* and *C₄* plants. *Planta* 237, 481–495. doi: 10.1007/s00425-012-1754-3
- Wang, P., Khoshravesh, R., Karki, S., Tapia, R., Balahadia, C. P., Bandyopadhyay, A., et al. (2017b). Re-creation of a key step in the evolutionary switch from *C₃* to *C₄* leaf anatomy. *Curr. Biol.* 27, 3278–3287. doi: 10.1016/j.cub.2017.09.040
- Wang, H., Seo, J. K., Gao, S., Cui, X., and Jin, H. (2017a). Silencing of *AtRAP*, a target gene of a bacteria-induced small RNA, triggers antibacterial defense responses through activation of LSU2 and down-regulation of *GLK1*. *New Phytol.* 215, 1144–1155. doi: 10.1111/nph.14654
- Wang, L., Tang, X., Zhang, S., Xie, X., Li, M., Liu, Y., et al. (2022). Tea *GOLDEN2-LIKE* genes enhance catechin biosynthesis through activating R2R3-MYB transcription factor. *Hortic. Res.* 9, 117. doi: 10.1093/hr/uhac117
- Wang, H., Zhang, D., Chen, M., Meng, X., Bai, S., Xin, P., et al. (2024). Genome editing of 3' UTR-embedded inhibitory region enables generation of gene knock-up alleles in plants. *Plant Commun.* 5, 100745. doi: 10.1016/j.xplc.2023.100745
- Waters, M. T., Wang, P., Korkaric, M., Capper, R. G., Saunders, N. J., and Langdale, J. A. (2009). GLK transcription factors coordinate expression of the photosynthetic apparatus in *Arabidopsis*. *Plant Cell.* 21, 1109–1128. doi: 10.1105/tpc.108.065250
- Wu, R., Guo, L., Guo, Y., Ma, L., Xu, K., Zhang, B., et al. (2023). The *G2-Like* gene family in *Populus trichocarpa*: identification, evolution and expression profiles. *BMC Genom. Data.* 24, 37. doi: 10.1186/s12863-023-01138-1
- Wu, R., Guo, L., Wang, R., Zhang, Q., and Yao, H. (2022). Genome-wide identification and characterization of G2-Like transcription factor genes in moso bamboo (*Phyllostachys edulis*). *Molecules* 27, 5491. doi: 10.3390/molecules27175491
- Xiong, B., Gong, Y., Li, Q., Li, L., Mao, H., Liao, L., et al. (2022). Genome-wide analysis of the *GLK* gene family and the expression under different growth stages and dark stress in sweet orange (*Citrus sinensis*). *Horticulturae* 8, 1076. doi: 10.3390/horticulturae8111076
- Xu, Z., Zhou, Z., Cheng, Z., Zhou, Y., Wang, F., Li, M., et al. (2023). A transcription factor *ZmGLK36* confers broad resistance to maize rough dwarf disease in cereal crops. *Nat. Plants.* 9, 1720–1733. doi: 10.1038/s41477-023-01514-w
- Yang, Z., Bai, T., Zhiguo, E., Niu, B., and Chen, C. (2024). OsNF-YB7 inactivates OsGLK1 to inhibit chlorophyll biosynthesis in rice embryo. *bioRxiv*. doi: 10.1101/2024.02.05.578907
- Yang, S., Wang, X., Yan, W., Zhang, Y., Song, P., Guo, Y., et al. (2023). *Melon yellow-green plant (Cmygp)* encodes a Golden2-like transcription factor regulating chlorophyll synthesis and chloroplast development. *Theor. Appl. Genet.* 136, 66. doi: 10.1007/s00122-023-04343-9
- Yasumura, Y., Moylan, E. C., and Langdale, J. A. (2005). A conserved transcription factor mediates nuclear control of organelle biogenesis in anciently diverged land plants. *Plant Cell.* 17, 1894–1907. doi: 10.1105/tpc.105.033191
- Yeh, S. Y., Lin, H. H., Chang, Y. M., Chang, Y. L., Chang, C. K., Huang, Y. C., et al. (2022). Maize Golden2-like transcription factors boost rice chloroplast development, photosynthesis, and grain yield. *Plant Physiol.* 188, 442–459. doi: 10.1093/plphys/kiab511
- Yelina, N. E., Frangedakis, E., Wang, Z., Schreier, T. B., Rever, J., Tomaselli, M., et al. (2024). Streamlined regulation of chloroplast development in the liverwort *Marchantia polymorpha*. *bioRxiv*. doi: 10.1101/2023.01.23.525199
- Ying, J., Wang, Y., Xu, L., Yao, S., Wang, K., Dong, J., et al. (2023). RsGLK2.1-RsNF-YA9a module positively regulates the chlorophyll biosynthesis by activating *RsHEMA2* in green taproot of radish. *Plant Sci.* 334, 111768. doi: 10.1016/j.plantsci.2023.111768
- Yuan, Y., Mei, L., Wu, M., Wei, W., Shan, W., Gong, Z., et al. (2018). SIARF10, an auxin response factor, is involved in chlorophyll and sugar accumulation during tomato fruit development. *J. Exp. Bot.* 69, 5507–5518. doi: 10.1093/jxb/ery328
- Zeng, X., Ye, L., Zhang, R., and Wang, P. (2023). Transcription factor GLK2 regulates key steps of anthocyanin biosynthesis to antagonize photo-oxidative stress during greening of *Arabidopsis* seedlings. *bioRxiv*. doi: 10.1101/2023.03.10.532066
- Zhang, H., Ji, Y., Jing, Y., Li, L., Chen, Y., Wang, R., et al. (2022a). *Arabidopsis* SIGMA FACTOR BINDING PROTEIN1 (SIB1) and SIB2 inhibit WRKY75 function in abscisic acid-mediated leaf senescence and seed germination. *J. Exp. Bot.* 73, 182–196. doi: 10.1093/jxb/erab391
- Zhang, Q. W., Mao, Y. Y., Zhao, Z. K., Hu, X., Hu, R., Yin, N. W., et al. (2024a). A Golden2-like transcription factor, BnGLK1a, improves chloroplast development, photosynthesis, and seed weight in rapeseed. *J. Integr. Agric.* 23, 1481–1493. doi: 10.1016/j.jia.2023.06.020
- Zhang, L., Qian, J., Han, Y., Jia, Y., Kuang, H., and Chen, J. (2022b). Alternative splicing triggered by the insertion of a CACTA transposon attenuates *LsGLK* and leads to the development of pale-green leaves in lettuce. *Plant J.* 109, 182–195. doi: 10.1111/tip.15563
- Zhang, D., Tan, W., Yang, F., Han, Q., Deng, X., Guo, H., et al. (2021). A BIN2-GLK1 signaling module integrates brassinosteroid and light signaling to repress chloroplast development in the dark. *Dev. Cell.* 56, 310–324. doi: 10.1016/j.devcel.2020.12.001
- Zhang, T., Zhang, R., Zeng, X. Y., Lee, S., Ye, L. H., Tian, S. L., et al. (2024b). GLK transcription factors accompany ELONGATED HYPOCOTYL5 to orchestrate light-induced seedling development in *Arabidopsis*. *Plant Physiol.* 194, 2400–2421. doi: 10.1093/plphys/kiac002
- Zheng, S. Y., Dong, J. F., Lu, J. Q., Li, J., Jiang, D. G., Yu, H. P., et al. (2022). A cytosolic pentatricopeptide repeat protein is essential for tapetal plastid development by regulating *OsGLK1* transcript levels in rice. *New Phytol.* 234, 1678–1695. doi: 10.1111/nph.18105
- Zhu, X., Zhang, L., Kuang, C., Guo, Y., Huang, C., Deng, L., et al. (2018). Important photosynthetic contribution of silique wall to seed yield-related traits in *Arabidopsis thaliana*. *Photosynth. Res.* 137, 493–501. doi: 10.1007/s11120-018-0532-x



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Tangren Cheng,
Beijing Forestry University, China
Yun-peng Du,
Beijing Academy of Agricultural and Forestry
Sciences, China

*CORRESPONDENCE

Jie Dong

✉ jiedong@nuau.edu.cn

Daidi Che

✉ daidiche@neau.edu.cn

RECEIVED 08 July 2024

ACCEPTED 13 August 2024

PUBLISHED 03 September 2024

CITATION

Duan L, Hou Z, Zhang W, Liang S, Huangfu M,
Zhang J, Yang T, Dong J and Che D (2024)
Genome-wide analysis of the *WOX* gene
family and function exploration of
RhWOX331 in rose (*R. 'The Fairy'*).
Front. Plant Sci. 15:1461322.
doi: 10.3389/fpls.2024.1461322

COPYRIGHT

© 2024 Duan, Hou, Zhang, Liang, Huangfu,
Zhang, Yang, Dong and Che. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/)
(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.

Genome-wide analysis of the *WOX* gene family and function exploration of *RhWOX331* in rose (*R. 'The Fairy'*)

Lian Duan^{1,2}, Zhihui Hou^{1,2}, Wuhua Zhang^{1,2}, Shuang Liang^{1,2},
Minge Huangfu^{1,2}, Jinzhu Zhang^{1,2}, Tao Yang^{1,2},
Jie Dong^{1,2*} and Daidi Che^{1,2*}

¹College of Horticulture and Landscape Architecture, Northeast Agricultural University, Harbin, China,

²Key Laboratory of Cold Region Landscape Plants and Applications, Harbin, China

*WOX*s are a class of plant-specific transcription factors that play key roles in plant growth and stress responses. However, the mechanism by which *WOX*s influence adventitious root development in *Rosa hybrida* remains unclear. In this study, *RcWOX* gene family in rose was identified and phylogenetically analyzed using bioinformatics analysis. A total of 381 *RcWOX* gene members were localized on seven chromosomes except of nine members. The main *cis*-acting elements involved in hormonal, light, developmental, and abiotic stress responses were identified in the promoters of *RcWOX* genes, suggesting their regulation by these signals. Nine *RhWOX* genes had significant different expression during rooting process of rose. *RhWOX331*, *RhWOX308*, *RhWOX318* were positive with the formation of rose roots. *RhWOX331* was positively involved in the formation of adventitious root primordia, which gene coding a transcription factor localized in the nucleus. The HOX conserved domain in the protein contributed to the self-activating activity of *RhWOX331*. We obtained genetically modified *Arabidopsis* to validate the function of *RhWOX331*. Overexpression of *RhWOX331* gene alleviated the inhibition of root length of *A. thaliana* primary roots by high concentration of IBA and NPA, and significantly increased the number of lateral roots on the primary roots, as well as the height of *A. thaliana* plants. Additionally, *RhWOX331* promoted adventitious root formation in *A. thaliana* and mitigated hormonal inhibition by exogenous 6-BA, NPA, and GA₃. The *RhWOX331* promoter contained *cis*-acting elements such as ABRE, Box 4 and CGTCA-motif et.al. GUS activity analysis showed that the gene acted at the cotyledon attachment site. Taken together, these studies identified a significant expansion of the *RcWOX* gene family, inferred roles of certain branch members in adventitious root formation, elucidated the function of *RhWOX331* in adventitious root initiation, and laid the foundation for further research on the function of *WOX* gene family in roses.

KEYWORDS

Rosa hybrida, *WOX* gene family, *RhWOX331*, adventitious roots, gene function analysis

1 Introduction

In agriculture, forestry and horticulture, plant organ regeneration was often utilized in cuttings propagation practices to obtain a large number of plants that retained the parent's good traits (De Klerk et al., 1999). For woody plants, the incidence of adventitious roots (ARs) during the propagation of cuttings determined the survival and efficiency of propagation of the species. ARs can be initiated from column sheath cells in the hypocotyl, thin-walled cells in the phloem or xylem, young secondary phloem cells, or cells of the inter bundle formation layer close to phloem cells (Bellini et al., 2014). The formation of adventitious roots was regulated by a combination of external environment, endogenous substances, and other factors, including light, water, spike age, and phytohormones (Bannoud and Bellini, 2021).

The WUSCHEL (WUS) homeobox transcription factor was a plant-specific transcription factor with a conserved “helix-loop-helix-turn-helix” motif comprising 60–66 amino acid residues (van der Graaff et al., 2009). During the phylogenetic process of higher plants, the WOX genes had evolved into three major classical clades: the modern/WUS clade, the ancient clade, and the intermediate clade. The modern evolutionary clade included WUS, WOX1~7, totaling 8 members, and the intermediate clade included 4 members, WOX8, WOX9, WOX11 and WOX12. The ancient clade members contained three genes, WOX10, WOX13 and WOX14 (Liu and Xu, 2018). Studies of the WOX gene family in *Arabidopsis thaliana* (Ohmori et al., 2013), *Populus trichocarpa* (Shuang et al., 2019), and *Picea abies* (Palovaara et al., 2010) had revealed that members of WOX gene family in each clade interacted with hormones to regulate plant growth and development processes. The WOX gene family played crucial regulatory roles during key stages of plant development such as embryo formation, stem cell maintenance, and organogenesis (Tanaka et al., 2015; Zhang et al., 2017), which were mediated by promoting cell division or inhibiting premature cell differentiation (Laux et al., 1996). These regulatory effects were likely achieved through interactions between WOX genes and hormones.

In modern clade, *AtWUS* regulated anther and ovule development (Reiser et al., 1995), and it also interacted with *CLAVATA3* (*CLV3*) to maintain the balance between proliferation and differentiation of stem tip meristems (Laux et al., 1996). *CsWUS* overexpression increased the number of sepals, petals and carpels in *Cucumis sativus* (Che et al., 2020). *AtWOX2* gene was expressed mainly in the apical cells of early embryonic development and regulated embryo formation (Liu and Xu, 2018). Overexpression of *AtWOX4* gene promoted radial growth of primary roots (Zhang et al., 2019). Among the genes in the intermediate clade, *AtWOX8* and *AtWOX9* were co-expressed in the pituitary cells of the embryo, promoted embryo development, and also functioned to maintain cell proliferation in the apical and root tip meristematic tissues (Liu and Xu, 2018). Overexpressing *PeWOX11a* or *PeWOX11b* in poplar not only enhanced adventitious root formation on the plugs, but also induced ectopic rooting in the aboveground part of transgenic poplar (Li et al., 2018). *OsWOX11* gene was expressed in the region of proliferative root tip cells and regulated the emergence and growth of

crown roots in rice (Cheng et al., 2014). There were fewer members in the ancient clade, among which *AtWOX13* functioned in early stages of root development and in organs with high proliferation, *AtWOX14* gene expressed in *A. thaliana* primary roots, lateral root primordia, and floral organs, and inhibited cell differentiation (Deyhle et al., 2007). Expression of *SkWOX13B* in stone pine plants was closely related to root organogenesis (Ge et al., 2016).

As the premier among the world's top four cut flowers, *Rosa hybrida* exhibits exceptionally high commercial value and possesses a unique cultural significance. In the genus *Rosa*, the ability to generate adventitious roots directly influences cutting survival and is a decisive factor in the garden application of *Rosa* species. WOX genes also regulated the growth and development of *Rosa* genus. The *RcaWOX1* gene from *Rosa canina* was induced by auxin and expressed at the early stage of healing tissue formation, overexpressing this gene increased the number of lateral roots and induced the up-regulated expression of *AtPIN1* and *AtPIN7* in *A. thaliana* (Gao et al., 2014). Overexpression of *RcWUS* induced the transformation of parenchyma cells in the root cortex into meristematic tissue cells, leading to the ectopic occurrence of adventitious shoots at the root tip (Jiang et al., 2012). The rooting ability of rose was influenced by factors such as genotype, lignification level of the cuttings, hormones, and environmental conditions. However, the molecular mechanisms underlying rose rooting remained unclear. This study provided a comprehensive overview of the WOX gene family in roses, investigating the expression patterns and functions of *RhWOX331* in adventitious rooting. It established a solid theoretical basis for further research on *RhWOX* genes involved in organogenesis in roses.

2 Materials and methods

2.1 Identification and phylogenetic analysis of WOX gene family members in rose

Genome and protein sequences of *Rosa chinensis*, *Rosa rugosa* and *Rosa multiflora* were obtained from the Rosaceae Genome Database (GDR) (Raymond et al., 2018; Jung et al., 2019). All WOXs in rose were identified using the Pfam protein family database (Mistry et al., 2021) by downloading the Hidden Markov Model (HMM) file for the WOX structural domain (PF00046) and setting a threshold of $1e^{-5}$. The core sequences of *RcWOXs* were verified using the SMART program and conserved domain database (CDD) (Wang et al., 2023). The protparam tool from the ExPASy website (<https://web.expasy.org/protparam/>) was used to predict basic characteristics (amino acid length, amino acid composition, isoelectric point, etc.) of the obtained WOX family members (Walker, 2005). Each *RcWOX* gene family member was named according to their position on the chromosome using TBtools II (Chen et al., 2023). The sequences of WOXs in *A. thaliana* (Lamesch et al., 2012), *Nicotiana tabacum*, and *Populus trichocarpa* were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). MEGA11 (Tamura et al., 2021) was used to perform multiple sequence comparisons under default

parameters, and a phylogenetic tree was constructed using the Neighbor-Joining method (bootstrap: 1000).

2.2 Analysis of WOXs of rose structure and conserved motifs

Multiple sequence comparisons were performed using ClustalW in MEGA11 under default parameters to further analyze the characteristic structural domains of RcWOX proteins and manually adjust the amino acid sequences. GSDS (Hu et al., 2015) was used for exon-intron structure visualization of RcWOX genes. RcWOXs protein motifs were analyzed using MEME (Bailey et al., 2009) under the parameter maximum motif number of 20.

2.3 Chromosomal localization, collinearity analysis and *cis*-acting element prediction of the RcWOX

Localization of all RcWOX genes to rose chromosomes based on physical location information using TBtools II (Chen et al., 2023). Utilizing TBtools II for collinearity analysis of the WOX gene family in rose with the WOX gene families of *A. thaliana* and *P. trichocarpa*. Promoter *cis*-acting regulatory elements were analyzed in the 2 Kb region upstream of the rose WOXs using PlantCARE (Lescot et al., 2002), and WOX gene family was visualized by TBtools II.

2.4 Plant materials and growth conditions

R. 'The Fairy' and *Nicotiana benthamiana* were grown in the Northeast Agricultural University (Harbin City, Heilongjiang Province, China) under a 16 h light/8 h dark at 25°C cycle. *A. thaliana* was grown under 14 h light/10 h dark conditions with a temperature range of 22–23°C and relative humidity between 40–60%.

2.5 Quantitative real-time PCR

Total RNA of plants leaves and roots was isolated with the FastPure Universal Plant Total RNA Isolation Kit (Vazyme Biotech Co., Ltd., Nanjing, China), and transcribed into cDNA using the HiScript III 1st Strand cDNA Synthesis Kit (+gDNA wiper) (Vazyme Biotech Co., Ltd., Nanjing, China). HiScript II QRT SuperMix for qPCR (Vazyme Biotech Co., Ltd., Nanjing, China) was used for qPCR. The determination of gene expression levels refers to previous research descriptions (Dong et al., 2021). The $2^{-\Delta\Delta CT}$ quantification method (Schmittgen and Livak, 2008) was used to calculate the relative expression levels. *RhActin* (Fan et al., 2023) were selected as reference genes in *Rosa hybrida*. All experiments were conducted with three biological replicates, each containing three technical repeats. Define a total of 8 stages from US to CS7 based on the cutting time of cuttings. US: 0 d; CS1: 15 min;

CS2: 1 d; CS3: 3 d; CS4: 5 d; CS5: 10 d; CS6: 15 d; CS7: 20 d. Primers used for RT-qPCR were listed in Supplementary Table S2.

2.6 Subcellular localization of RhWOX331

The full-length *RhWOX331* gene, lacking a stop codon, was inserted into *KpnI* and *BamHI* sites (Takara, Beijing, China) of the pGAMBIA1300-sGFP vector using the pEASY[®]-Basic Seamless Cloning and Assembly Kit (TransGen Biotech, Beijing, China). The constructed vector pGAMBIA1300-*RhWOX331*-sGFP was transformed into *Agrobacterium tumefaciens* GV3101 (WeiDi Biotechnology, Shanghai, China), and subcellular localization was performed according to the previous research (Li et al., 2023). The infection solution (200 μ M acetosyringone (AS), 10 mM 2-morpholinoethanesulphonic acid (MES), and 10 mmol/l $MgCl_2$) containing either pGAMBIA1300-*RhWOX331*-sGFP or pGAMBIA1300-sGFP were injected into the subepidermal cells of 4-week-old *Nicotiana benthamiana* leaves. After 2 days of dark incubation at 23 °C, the subcellular localization of RhWOX331 was visualized and photographed using a laser-scanning confocal microscope (FV3000, Olympus, Japan) at 488 nm.

2.7 Yeast self-activation analysis of RhWOX331

The full-length *RhWOX331* gene was inserted into *NdeI* and *EcoRI* (Takara, Beijing, China) sites of the pGBKT7 vector using the pEASY[®]-Basic Seamless Cloning and Assembly Kit (TransGen Biotech, Beijing, China). The pGADT7-T+pGBKT7-p53 (positive control), pGADT7-T+pGBKT7-lam (negative control), and pGBKT7-WOX331-1, pGBKT7-WOX331-2, pGBKT7-WOX331-3 plasmids were transformed into Y2HGold yeast competent cells (WeiDi Biotechnology, Shanghai, China). After 2 days of cultivation at 28°C, yeast colonies were selected and cultured in SD/-Trp/-Leu liquid yeast medium at 28°C and 200 rpm. Centrifuge yeast at 4000 rpm for 1 minute to collect the yeast cells. The Y2HGold yeast containing the recombinant plasmid was resuspended in sterile water until its OD600 reached 0.2. The suspended culture was diluted to 1X, 10X, and 100X concentrations. The positive control and negative control diluted yeast solution was placed on SD/-Trp/-Leu/-His/-Ade/X- α -gal solid medium, the diluted yeast solution transforming pGBKT7-WOX331-1, pGBKT7-WOX331-2, pGBKT7-WOX331-3 was placed on SD/-Trp/-His/X- α -gal solid medium and cultured at 28°C. After 36–48 h of incubation, the self-activating activity of *RhWOX331* was assessed based on the blue coloration of the yeast.

2.8 Genetic transformation and identification of transgenic RhWOX331 in A. thaliana

pGAMBIA1300-*RhWOX331*-sGFP was transformed in *A. thaliana* with floral dip transformation method (Bent, 2006).

Seeds of *A. thaliana* were collected and sown, until obtaining T3 generation plants. Transgenic *A. thaliana* were identified by PCR using WOX331F and WOX331R as primers (Supplementary Table S2). The seeds of transgenic *A. thaliana* were sterilized and sown in 1/2 MS medium (20 g/L sucrose + 8 g/L agar), and different hormones were added to the medium according to different treatments: CK: no hormone; IBA: 0.25 mg/L IBA; 6-BA: 0.5 mg/L 6-BA; GA₃: 1 mg/L GA₃; NPA: 10 μM NPA. The phenotypes of the primary roots of *A. thaliana* were determined after 14 days. At 14d, the main roots were removed and transferred to B5 medium (30 g/L sucrose + 8 g/L agar), and different hormones were added to the medium according to different treatments (hormone concentration as above), and the phenotypic changes of adventitious roots were observed.

2.9 Analysis of the GUS activity of *RhWOX331* promoter

The 2113 bp sequence upstream of the start codon of *RhWOX331* was divided into three segments. *pWOX331* replaced 35S in PBI121 and construct *pWOX331-1::GUS*, *pWOX331-2::GUS* and *pWOX331-3::GUS* vectors with *Bam*HI and *Hind*III restriction site. Primers were listed in Supplementary Table S2. Transgenic *A. thaliana* overexpressing *pWOX331-1::GUS*, *pWOX331-2::GUS* and *pWOX331-3::GUS* were immersed in GUS staining solution (Coolaber, Beijing, China) and kept warm at 37°C for 1 h. Using 70% ethanol for decolorization 2~3 times, and the material was observed under the *in vitro* microscope (Olympus SZX2-ILLTQ). P1, P2, and P3 represent *A. thaliana* transformed with *pWOX331-1::GUS*, *pWOX331-2::GUS*, and *pWOX331-3::GUS*, respectively. GUS activity was assessed using the previous method (Koo et al., 2007).

2.10 Statistical analyses

Statistical analyses were conducted with IBM SPSS v25.0 (SPSS Inc., Chicago, IL, USA). Least Significant Difference (LSD) test was performed in order to compare the statistical validity of data. Significance was set at $p < 0.05$. Three biological replicates were used for each assay. TBtoolsII software was used to create the conserved domains, motifs, gene structure. GraphPad Prism 8.0.0 (GraphPad Software San Diego, California USA) were used to plot graphs.

3 Results

3.1 Identification of WOXs in rose

The 381 members of the rose WOX gene family were finally identified in the whole rose genome, and they were named *RcWOX1-RcWOX381* based on their positions on the chromosome (Supplementary Table S1). The physicochemical properties of the 381 WOX genes revealed that the number of amino acids ranged from 81 to 400, and the theoretical isoelectric

points ranged from 4.56 to 10.55, with 87.9% of them having isoelectric points lower than 7, indicating that they were mostly acidic proteins. The instability coefficients ranged from 36.55% to 86.47%, with 2.1% of the members having instability coefficients lower than 40%, and most of the WOX proteins were unstable proteins. The relative molecular mass of *RcWOX335* was the largest at 44.66 KDa, and the relative molecular mass of *RcWOX275* was the smallest at 9.87 KDa.

3.2 Phylogenetic analysis

In order to explore the phylogenetic relationships of WOXs in rose and other model plants, a phylogenetic tree was constructed based on the sequences of 453 WOX proteins from rose (381), *P. trichocarpa* (26), *N. tabacum* (28) and *A. thaliana* (18) (Figure 1). The phylogenetic tree analysis showed that the 453 genes were clearly divided into eight clades: ancient clade, intermediate clade, modern/WUS clade, clade I, clade II, clade III, clade IV, clade V. Among these, the *RcWOXs* in the classical clades including ancient, intermediate, and modern/WUS clades were more closely related to *N. tabacum*, *A. thaliana*, and *P. trichocarpa* WOXs. On the contrary, WOX genes in rose belonging to clades I to V had no homologous genes with *P. trichocarpa*, *N. tabacum* and *A. thaliana* WOX genes. Ancient clade contained 2 genes in rose, 3 genes in *A. thaliana*, 6 genes in *P. trichocarpa*, 6 genes in *N. tabacum*. Intermediate clade contained 2 genes in rose, 7 genes in *A. thaliana*, 6 genes in *P. trichocarpa*, 6 genes in *N. tabacum*. Modern/WUS clade contained 13 genes in rose, 8 genes in *A. thaliana*, 14 genes in *P. trichocarpa*, 16 genes in *N. tabacum*. Clades I to V contained 364 members, all of which originated from rose. Clade V was the largest clade, containing 265 members. The results show that there are a large number of similar redundant genes in rose WOX gene family, and they are distantly homologous to the WOX family members of the ancient, intermediate, and modern/WUS clades.

Based on the chromosomal location information of *RcWOXs* in *R. chinensis*, the positions of 381 *RcWOX* members on the chromosomes were visualized and analyzed (Figure 2). *RcWOXs* were distributed on all seven chromosomes, with a total of 226 *RcWOX* genes on chromosome 2, 46 *RcWOX* genes on chromosome 3, 38 *RcWOX* genes on chromosome 7, 36 *RcWOX* genes on chromosome 1, 10 *RcWOX* genes on chromosome 5, 9 *RcWOX* genes on chromosome 6, 7 *RcWOX* genes on chromosome 4, and 9 *RcWOX* genes not localized on any chromosome. WOX genes were most densely distributed on chromosome 2.

Conserved domain analysis of 381 *RcWOX* family members revealed the presence of two conserved domains: Homeodomain superfamily and Homeobox (Supplementary Figure S1B). In order to study the structure of *RcWOXs*, a figure depicting the *RcWOX* structure was created (Supplementary Figure S1A), which showed that motifs 10 and 15 were present in all members of *RcWOXs* of classical clades. In contrast, the vast majority of the members in clades I to V contained motifs 1, 2 and 4. The gene structure figure also indicated that 86.8% of members in clades I to V and 29.4% of members in classical clades lacked UTRs (Supplementary Figure S1C). Analysis of the amino acid sequences of clade V revealed that

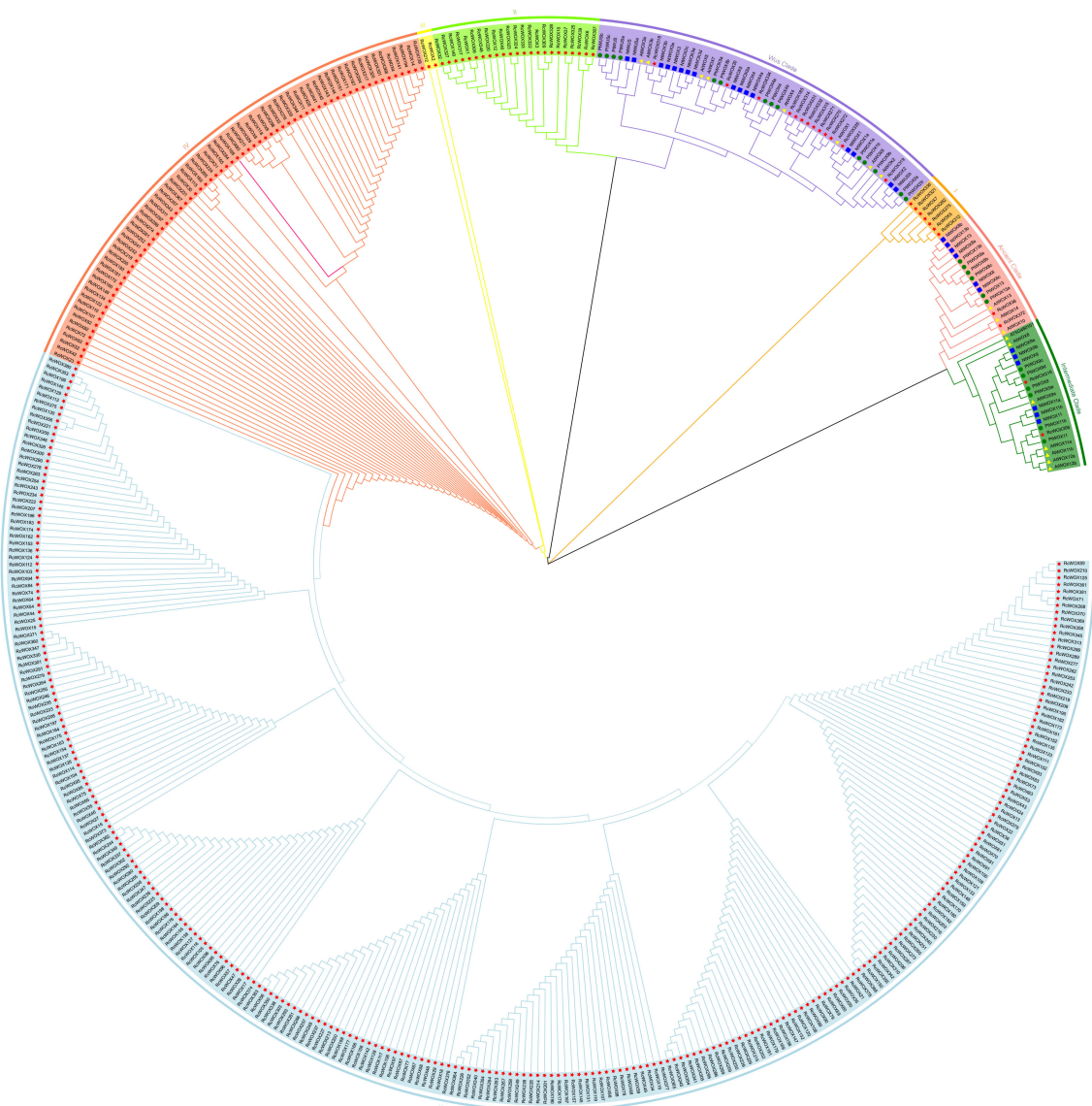


FIGURE 1

Phylogenetic analysis of the *RcWOX* gene family of *R. chinensis* and *N. tabacum*, *A. thaliana*, *P. trichocarpa*. Proteins from *R. chinensis*, *N. tabacum*, *A. thaliana* and *P. trichocarpa* were respectively denoted by the prefixes Rc, Nt, At, and Pt, respectively. They were divided into eight major phylogenetic clusters: ancient clade, intermediate clade, WUS clade, clade I, clade II, clade III, clade IV, and clade V. Each clade was indicated by different colors. Bootstrap:1000.

most of the proteins contained the amino acid domains SIMEQRGBYHQBIBTLPLFPMHGEDI LGNMKTTS EGGGGGYGG and G/DSHISLELSLNSYRDADMA, corresponding to motifs 2 and 4. For clade IV of *WOX*s in rose, most proteins contained the amino acid domains HQEITLMHGEDI and YGQIEDKNVFFWFQNLKA, which were absent in classical clades. These findings suggest significant differences in amino acid sequences, conserved domains, and intron distribution between *WOX* members of classical clades and clades I to V, implying potential functional distinctions.

The *cis*-acting elements within the upstream 2000bp of the initiation codon of 381 *WOX* genes in rose were involved in hormone, environment, growth and development (Supplementary Figure S2). Hormone-related *cis*-acting elements were salicylic acid-

induced (W-box), jasmonic acid signaling pathway (MYC), and gibberellin response element (P-box). *Cis*-acting elements involved in environment including light-responsive (G-box, Sp1, TGACG-motif, TCCC-motif) and trauma response (WUN-motif). MYB, GCN4-motif, Circadian clock belonged to growth and development-related *cis*-acting elements. It was observed that *WOX* genes within the same clade of the phylogenetic tree exhibited similar *cis*-acting elements. These findings suggest that *WOX* genes in rose may be regulated by a diverse array of phytohormones, biotic and abiotic stimuli, influencing plant growth and development.

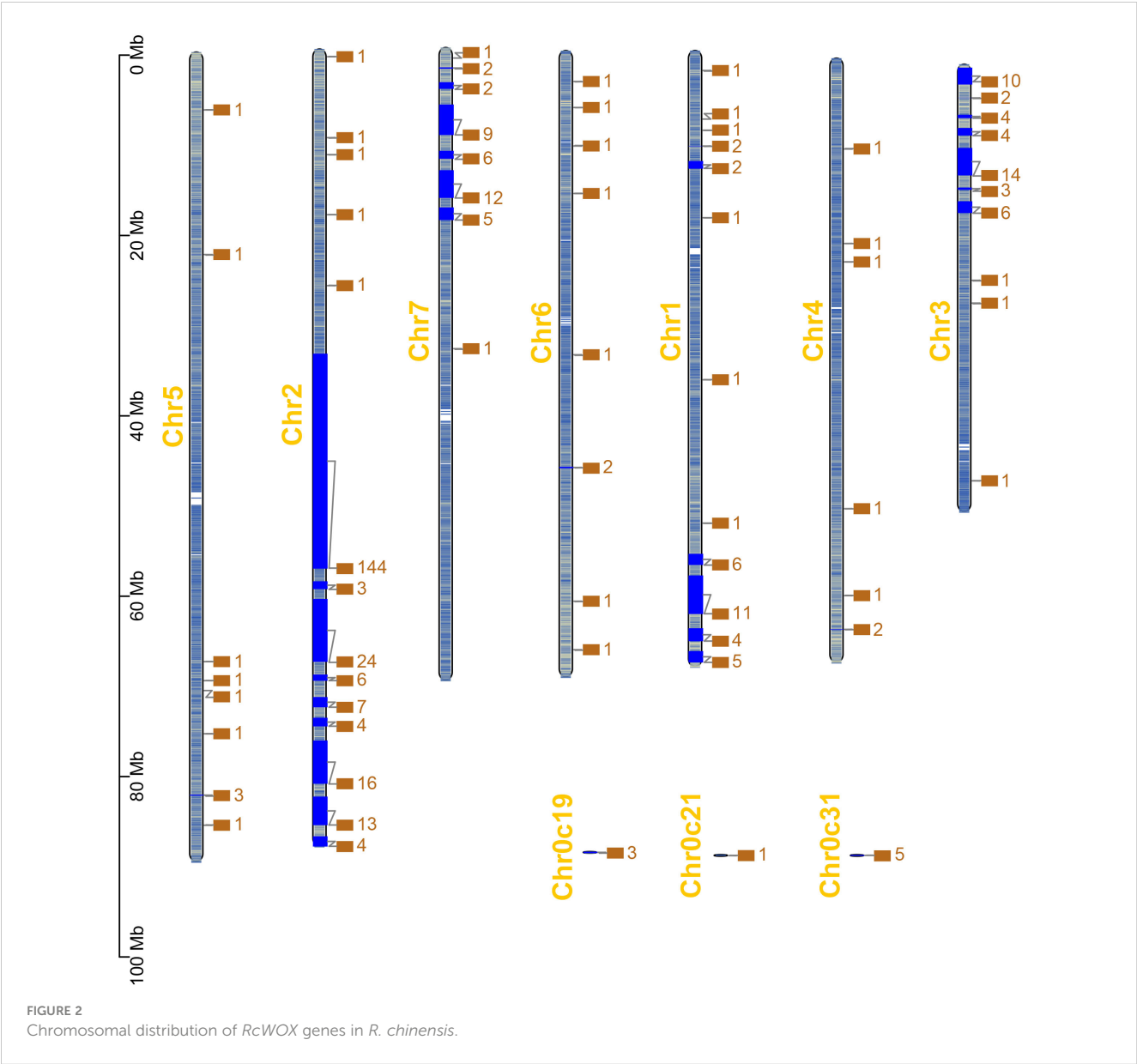
In order to further investigate the interspecific evolutionary relationship of *WOX*s, intergenic collinearity analysis was performed between roses and model plants, such as *A. thaliana*

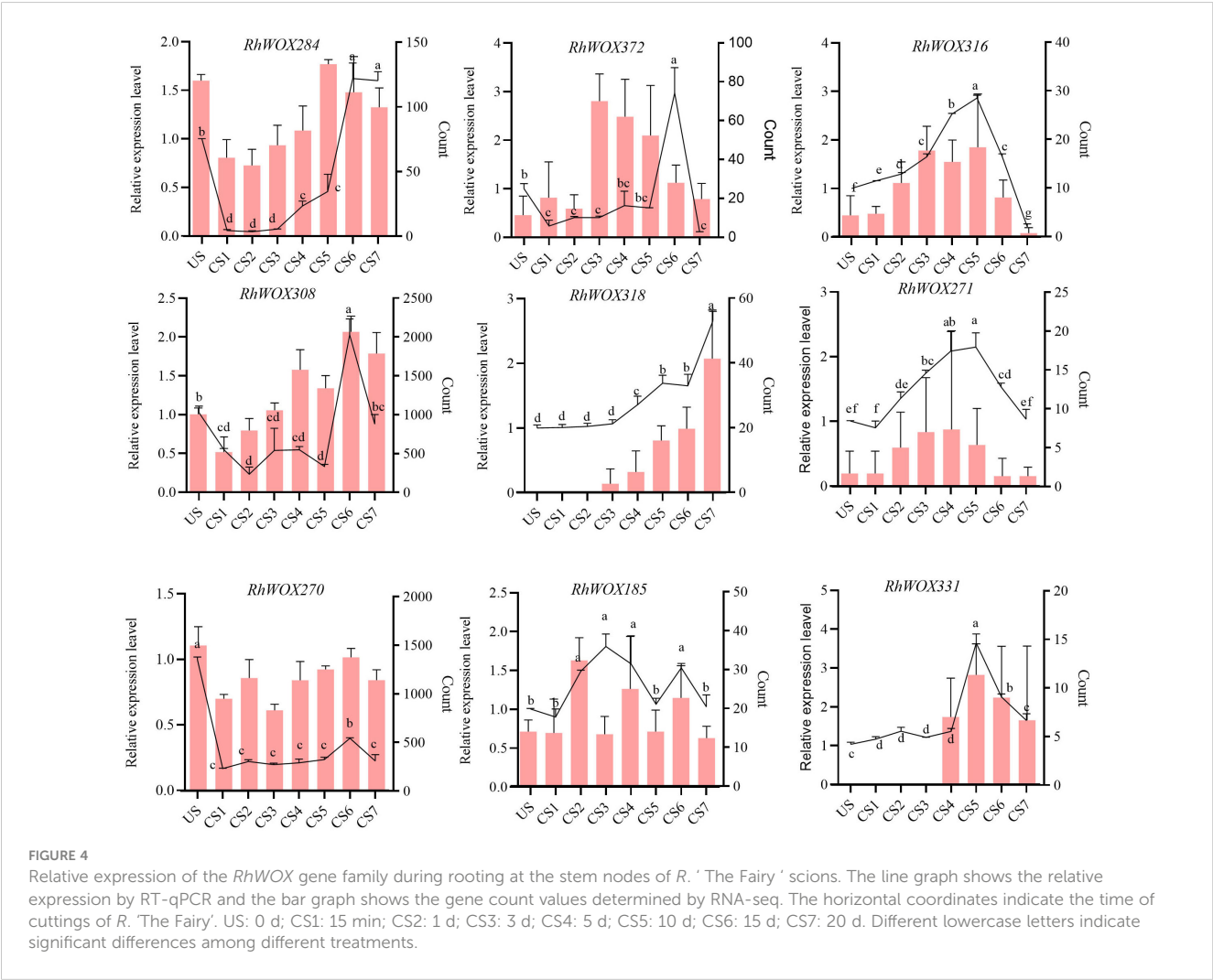
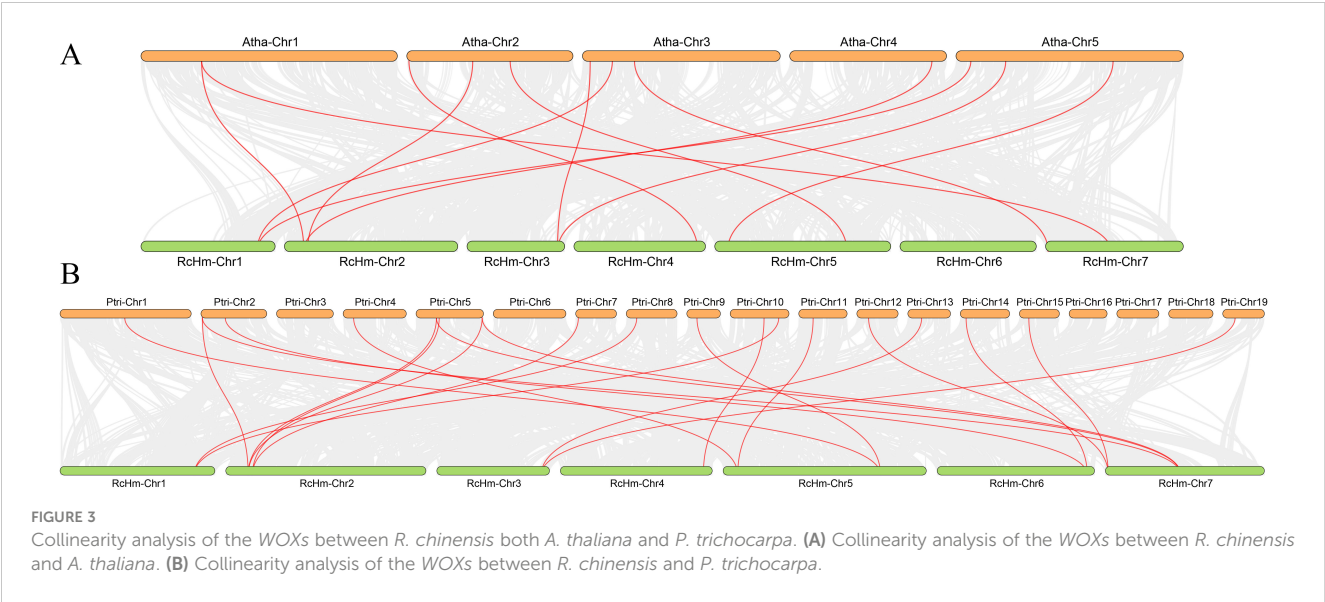
(Figure 3A) and *P. trichocarpa* (Figure 3B). It was found that there were 12 homologous gene pairs between 381 *WOX* genes of rose and 15 *AtWOX* genes of *A. thaliana*, and 21 homologous gene pairs between 381 *WOX* gene family members of rose and 26 *PtWOX* genes of *P. trichocarpa*. These results suggest a high number of homologous gene pairs between rose *WOX* genes and both *AtWOXs* and *PtWOXs*.

3.3 Analysis of *RhWOXs* expression patterns

The expression levels of *WOX* genes during adventitious root formation in roses indicated relatively higher expression levels in classical clades, with almost all members in clades I to IV showing no expression. Therefore, we selected *RhWOX* genes in ancient, intermediate, and modern/WUS clades. Combining the expression

data of *WOXs* transcripts during adventitious rooting process of *R. hybrida*, nine *RhWOXs* were finally identified (Figure 4). Expression analysis of *RhWOXs* gene family members during rooting of single-node spikes of rose showed that *RhWOX284* was down-regulated 15 minutes after cutting, and up-regulated during leaf production. *RhWOX372*, *RhWOX316* and *RhWOX271* were up-regulated in the mid-root stage of CS3~CS5, and down-regulated in the root elongation stage. *RhWOX308* and *RhWOX270* were initially down-regulated after pruning, and these genes were significantly up-regulated as the stem cells divided and root primordia formed. *RhWOX318* gradually activated during root tip formation, exhibiting peak activity during root elongation. *RhWOX185* showed significant up-regulation during CS2 stage. *RhWOX331* remained low until root primordium formation (CS1~CS4), exhibited significant up-regulation during CS4~CS5, and then down-regulated during the period of root tip formation and root elongation, showing strong correlation with root primordium





development. RT-qPCR data corroborated RNA sequencing results, with *RhWOX331* showing significant positive correlation with root primordium differentiation. Thus, we speculate that *RhWOX331* gene play a key role in the development of adventitious roots in *R. hybrida*.

3.4 Characterization of *RhWOX331*

The expression of *RhWOX331* showed tissue-specificity, with the highest expression in roots, followed by stems, and the lowest expression in flowers (Figure 5A). Exogenous application of IBA promotes adventitious root formation in roses, whereas NPA application suppresses it. By the 10th day of cutting, exogenous IBA significantly increased the expression of *RhWOX331* to 1.3 times that of the hormone-free control, while exogenous NPA significantly reduced *RhWOX331* expression to 0.8 times that of the hormone-free control (Figure 5B). IBA promoted the expression of *RhWOX331* continuously. In the absence of hormones, expression of *RhWOX331* in cuttings remained almost unchanged after 5 d of cultivation. The gene was up-regulated from 5 to 10 days

and then down-regulated from 15 to 20 days. After the application of exogenous IBA, *RhWOX331* showed upregulated expression as early as 5 d after culture initiation. At each time point thereafter, the expression level of this gene was significantly higher compared to the control without any hormone addition (Figure 5C).

Subcellular localization analysis revealed that in the control tobacco leaf cells, green fluorescence can be observed simultaneously in both the cell membrane and nucleus. In the *RhWOX331* group, only green fluorescence was observed within the nucleus, confirming the nuclear localization of *RhWOX331* (Figure 5D). To verify the transcriptional activation activity of *WOX331*, three segments of the *WOX331* gene were constructed into the *pGBKT7* vector (Figure 5E). On SD/-Ade/-His/-Leu/-Trp medium, the negative control yeast did not grow, while the positive control yeast grew and turned blue after adding X- α -gal. The yeast that transformed *pGBKT7-WOX331-1* did not grow, while the yeast that transformed *pGBKT7-WOX331-2* and *pGBKT7-WOX331-3*, which both containing the HOX domain grew normally and turned blue after adding X- α -gal (Figure 5F). This indicates that the transcription factor *RhWOX331* possesses self-activation activity, which may be attributed to the HOX domain spanning amino acids 87 to 807.

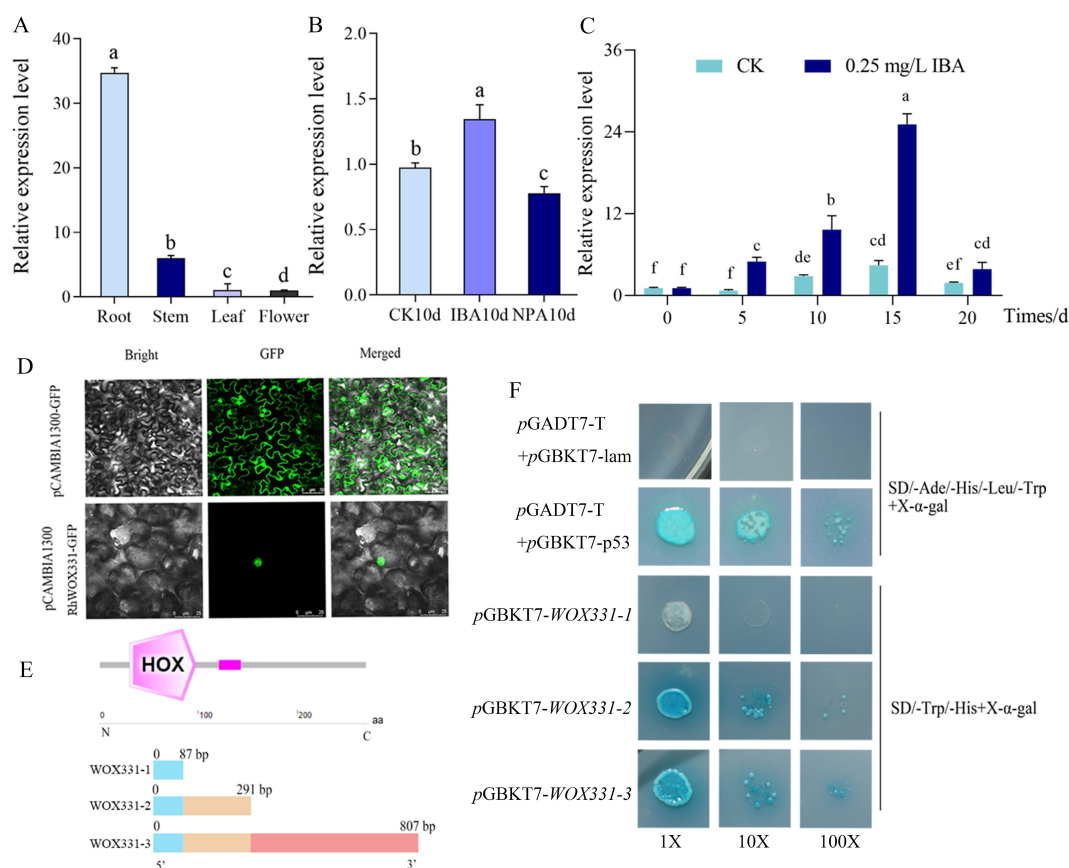


FIGURE 5

Characterization of *RhWOX331*. (A) Expression of *RhWOX331* in different tissues of *R. 'The Fairy'*. (B) Relative expression of *RhWOX331* gene in 0.25 mg/L IBA, 0.2 mg/L NPA-treated and untreated rose cuttings. (C) Trend of *RhWOX331* gene expression during adventitious root primordia formation in *R. 'The Fairy'* cuttings. (D) Subcellular localization of *RhWOX331*. pCambia1300-GFP empty vector as a control. (E) Schematic diagram of yeast self-activation vector construction for *RhWOX331*. (F) Verification of yeast self-activation of *RhWOX331*. pGADT7-T+pGBKT7-p53 served as the positive control, while pGADT7-T+pGBKT7-lam was utilized as the negative control. Different lowercase letters indicate significant differences among different treatments.

3.5 The effect of overexpression of *RhWOX331* on rooting and growth of *A. thaliana* seeds

RhWOX331-overexpressing *A. thaliana* lines were obtained and identified to investigate the influence of *RhWOX331* on root development (Supplementary Figure S3). There was no significant difference in root length between wild-type (WT) and *RhWOX331*-overexpressing *A. thaliana* plants on hormone-free medium or medium containing 0.5 mg/L 6-BA or 1 mg/L GA₃ (Figure 6B). The average root length of 14-day-old plants was approximately 6.78 cm in the CK, 0.74 cm in the 6-BA group and 4.6 cm in the GA₃ group. Interestingly, on medium containing 0.25 mg/L IBA and 10 μM NPA, *A. thaliana* growth was inhibited, showing differences in root length between WT and transgenic plants (Figures 6A, C, F). The root lengths

of WT plants were 1.34 cm and 1.91 cm, respectively. However, overexpression of *RhWOX331* alleviated the inhibitory effects of these high concentrations of exogenous hormones, resulting in primary root lengths of 3.08 cm and 2.86 cm, respectively. The number of lateral roots on the primary root had also significantly increased. Moreover, it was found that both the plant height and the height between the capsules and the rosette were increased after overexpressing *RhWOX331* (Figures 6G, H, I). It was different that the number of capsules did not increase (Figure 6J). These results indicate that overexpression of *RhWOX331* did not promote elongation of primary roots in *A. thaliana*, but enhanced lateral root formation. It also alleviated the inhibitory effects of high concentrations of auxin and auxin inhibitors on primary root elongation. Moreover, *RhWOX331* increased plant height by raising the height between the capsules and the rosettes rather than increasing the number of flowers.

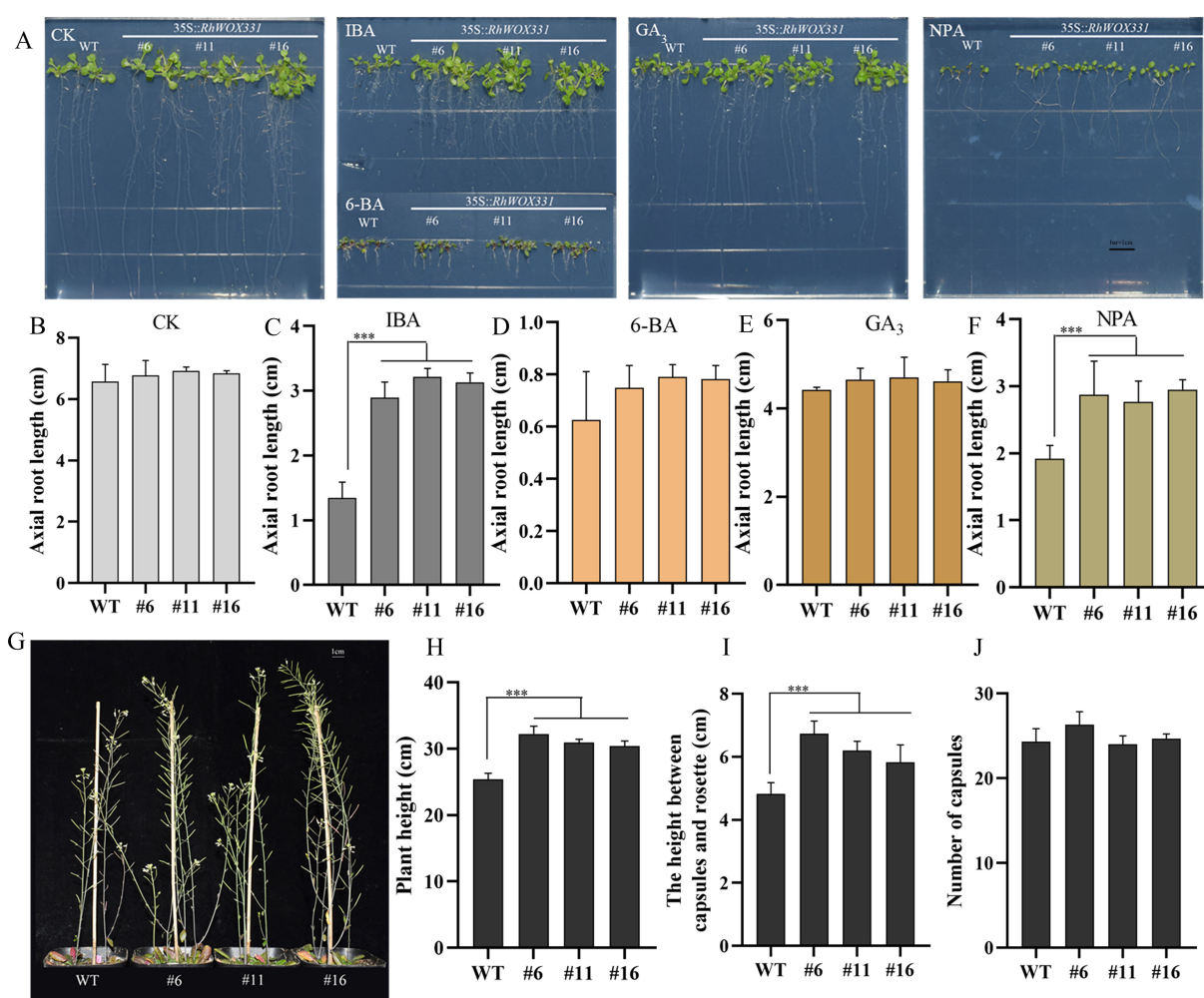


FIGURE 6

Growth of *A. thaliana* seeds overexpressing *RhWOX331*. (A) Primary root length of WT and *RhWOX331* overexpressing *A. thaliana* seeds after 14 days of cultivation in different mediums. CK) hormone-free medium; IBA) medium containing 0.25 mg/L IBA; 6-BA) medium containing 0.5 mg/L 6-BA; GA₃) medium containing 1 mg/L GA₃; NPA) medium containing 10 μM NPA. (B-F) Primary root length of WT and *RhWOX331* overexpressing *A. thaliana* seeds in different medium. Bar = 1 cm. (B) hormone-free medium; (C) medium containing 0.25 mg/L IBA; (D) medium containing 0.5 mg/L 6-BA; (E) medium containing 1 mg/L GA₃; (F) medium containing 10 μM NPA. (G) The phenotypes of mature WT and *RhWOX331* overexpressing *A. thaliana* plants. Bar = 1 cm. (H) The plant height of mature WT and *RhWOX331* overexpressing *A. thaliana* plants. (I) The height between the capsules and the rosette of mature WT and *RhWOX331* overexpressing *A. thaliana* plants. (J) the number of capsules of mature WT and *RhWOX331* overexpressing *A. thaliana* plants. The *** mark indicates significant difference between WT and transgenic lines.

3.6 The effect of overexpression of *RhWOX331* on the rooting of *A. thaliana* adventitious roots

The primary roots of 14-day-old *A. thaliana* were removed and cultivated on B5 medium containing different hormones. Adventitious root formation in *A. thaliana* was enhanced on hormone-free medium and medium containing 0.25 mg/L IBA. Overexpression lines exhibited earlier adventitious root emergence, with a greater number and longer lengths of adventitious roots compared to the WT (Figures 7A–C, G, H). The difference in the number of adventitious roots was particularly pronounced. On the medium containing 0.5 mg/L 6-BA, 1 mg/L GA₃, and 10 μM NPA, WT plants almost did not form roots after 10 days of culture, whereas *RhWOX331* overexpressing plants developed some adventitious roots (Figures 7A, D–F, I–K). In terms of both the

number and length of adventitious roots, *RhWOX331* overexpressing *A. thaliana* demonstrated a stronger rooting ability. These results suggest that overexpression of *RhWOX331* promotes adventitious root formation in *A. thaliana* and alleviates the inhibitory effects of some hormones on adventitious root development.

3.7 Analysis of GUS activity of *RhWOX331* promoter

By predicting the approximately 2000bp sequence upstream of the *RhWOX331* gene start codon, it was found that this sequence contained abundant *cis*-regulatory elements (Supplementary Figure S2), which may be one of the reasons that *RhWOX331* was regulated by many hormones, such as IBA. According to the position of the

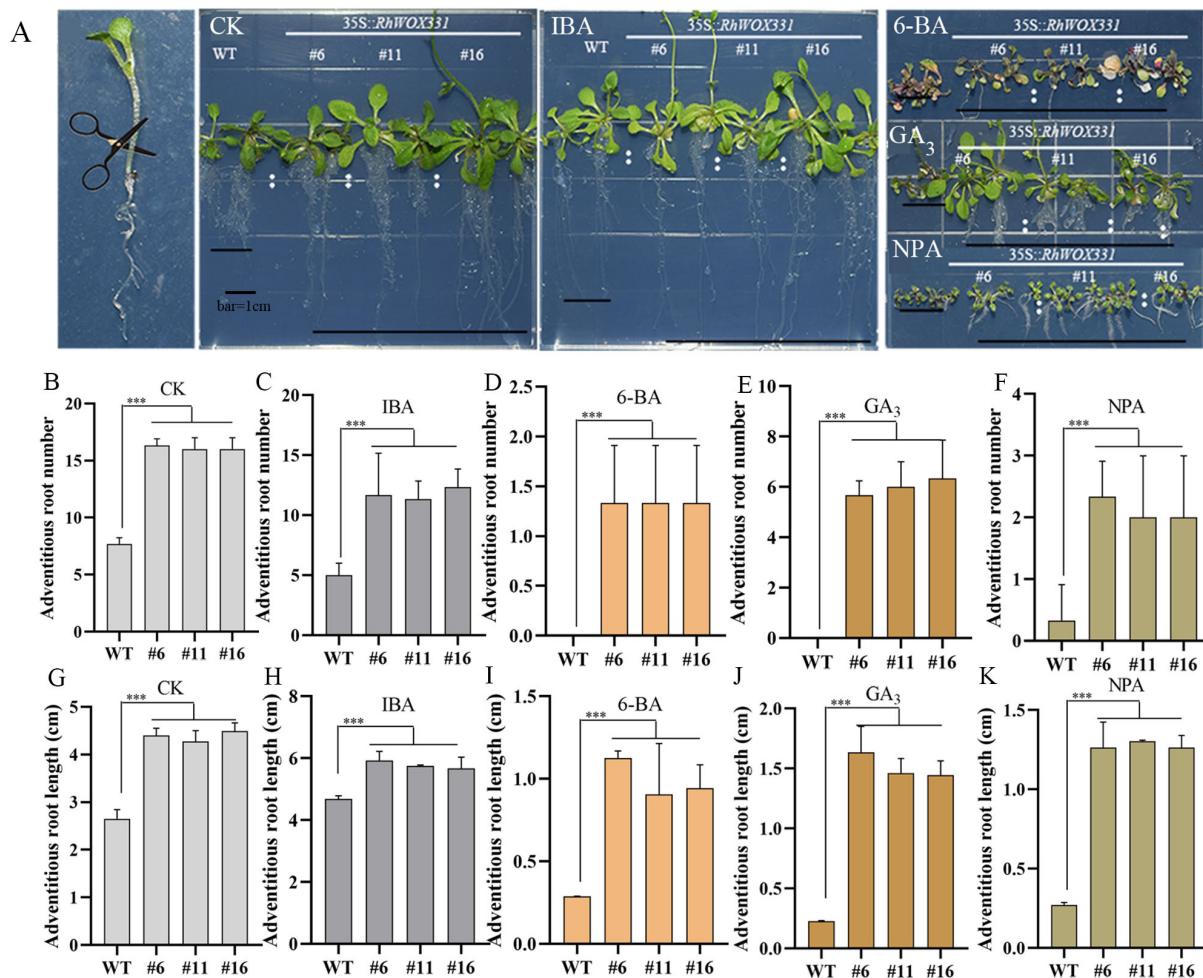


FIGURE 7

Growth of adventitious roots of WT and *RhWOX331* overexpressing *A. thaliana* in medium containing different hormones. (A) After removal of the primary root, WT and *RhWOX331* overexpressing *A. thaliana* developed adventitious roots on medium containing different hormones: CK) hormone-free medium; IBA) medium containing 0.25 mg/L IBA; 6-BA) medium containing 0.5 mg/L 6-BA; GA₃) medium containing 1 mg/L GA₃; NPA) medium containing 10 μM NPA. (B–F) The number of adventitious roots occurring in WT and overexpressed *RhWOX331* *A. thaliana* in different medium. Bar = 1 cm. (B) hormone-free medium; (C) medium containing 0.25 mg/L IBA; (D) medium containing 0.5 mg/L 6-BA; (E) medium containing 1 mg/L GA₃; (F) medium containing 10 μM NPA. (G–K) Length of adventitious roots occurring in WT and overexpressed *RhWOX331* *A. thaliana* in different media. (G) hormone-free medium; (H) medium containing 0.25 mg/L IBA; (I) medium containing 0.5 mg/L 6-BA; (J) medium containing 1 mg/L GA₃; (K) medium containing 10 μM NPA. The *** mark indicates significant difference between WT and transgenic lines.

TATA-box, the 2113bp sequence was divided into three segments (Figure 8A), *pWOX331-1::GUS*, *pWOX331-2::GUS*, and *pWOX331-3::GUS* vectors were constructed and transformed into *A. thaliana* (Figure 8B). P1, P2, and P3 represent *A. thaliana* transformed with *pWOX331-1::GUS*, *pWOX331-2::GUS*, and *pWOX331-3::GUS*, respectively. Observation of GUS staining in 7-day-old *A. thaliana* seedlings revealed no blue spots in the WT plants, while GUS signals were detected at the shoot apical meristem and cotyledonary node in overexpressing plants carrying *pWOX331-1::GUS* and *pWOX331-2::GUS* vectors (Figure 8C). GUS activity of *RhWOX331* promoter showed the same results (Figure 8D). Considering that no GUS signal was detected in transgenic *A. thaliana* after adventitious root formation, it was speculated that *WOX331* played a role before visible adventitious root formation. These results indicate that the promoter of *WOX331* is located between 731bp and 2113bp. In addition to regulating adventitious root formation, *RhWOX331* also plays a role in the growth point of *A. thaliana* cotyledons.

4 Discussion

4.1 The *RcWOX* gene family had undergone significant expansion in *Rosa chinensis*

A total of 381 *WOX* genes were identified in rose, a number significantly higher than that found in other species, including 18 in *A. thaliana*, 28 in *N. tabacum*, 26 in *P. trichocarpa* (Figure 1), 18 in *Eriobotrya japonica* (Yu et al., 2022) and 33 in *Glycine max* (Hao et al., 2019). The occurrence of more than 100 members in the *WOX* gene family was not unique to roses. Other species within the

Rosa genus which had published genomes also had a relatively large number of *WOX* genes. *Rosa multiflora* contained 170 *WOX* genes, and 105 *WOX* genes were identified in *Rosa rugosa*. The number of *WOX* genes in rose was substantially higher than that in other species, but in other Rosaceae species, the number of *WOX* genes was not particularly high. There were 9–14 *WOX* gene family members in *Pyrus bretschneideri* and other Rosaceae species (Cao et al., 2017). Lv identified *WOX* gene family members in nine *Prunus* species, ranging from 6 to 40 (Lv et al., 2023). The number of *WOX* genes in *R. chinensis*, *R. multiflora* and *R. rugosa* was also above average, suggesting that the large-scale expansion of the *WOX* gene family was a phenomenon specific to the genus *Rosa*. The genetic background of *R. chinensis* was relatively complex, and *Rosa multiflora* and *Rosa rugosa* might be involved in the breeding process of this species (Cui et al., 2022). The *WOX* gene may have replicated during this process. At present, there is no analysis on the *WOX* function of roses, and more genetic functional evidence is needed to determine the specific significance of this replication process. Whole genome duplications (WGD) are the primary driver of *WOX* family evolution (Cao et al., 2017). In *Bromeliaceae* plants, the CAM-related gene families had experienced accelerated expansion, supporting gene family evolution as a driver of CAM evolution (Groot Crego et al., 2024). Abubakar identified four segmental duplications and one tandem duplication of *WOX* gene family in *Boehmeria nivea* (Abubakar et al., 2023), which suggested that whole-genome duplication (WGD) had contributed to the expansion of the *WOX* gene family in *B. nivea*. During the Paleocene–Eocene boundary, Rosaceae underwent a WGD event, leading to extensive gene duplication (Xiang et al., 2017). The entire *Malus* genus experienced a WGD event, resulting in the duplication of several MADS-box genes potentially linked to pome formation during that period (Zhang et al., 2023). We hypothesize

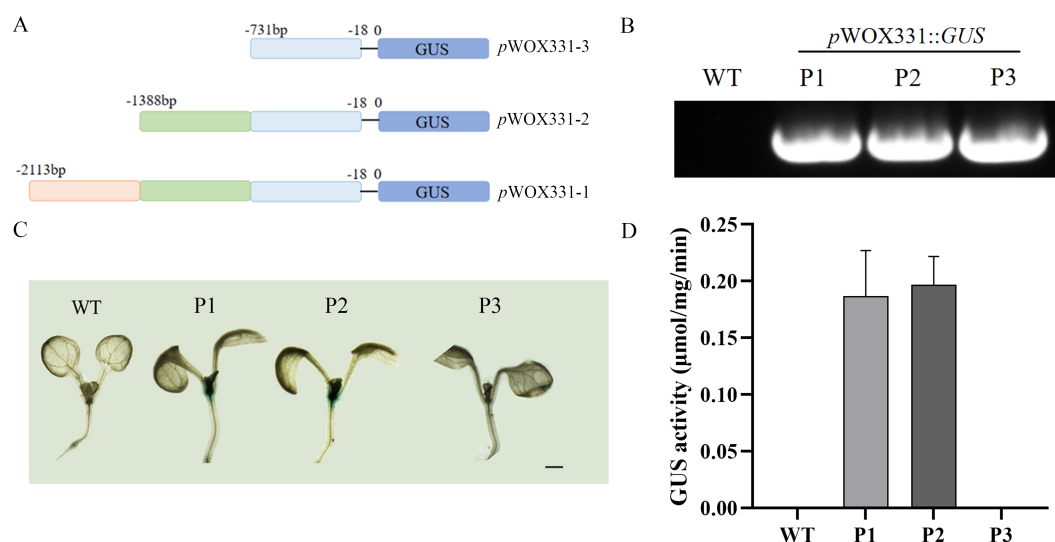


FIGURE 8

Analysis of GUS activity of *RhWOX331*. (A) Schematic diagram of *RhWOX331* GUS vector construction. (B) Identification of *A. thaliana* transformed with promoter of *RhWOX331*. P1, P2, and P3 represent *A. thaliana* transformed with *pWOX331-1::GUS*, *pWOX331-2::GUS* and *pWOX331-3::GUS*, respectively. (C) GUS staining of WT and *A. thaliana* overexpressing *pWOX331*. Bar = 1 mm. (D) GUS activity of WT and *A. thaliana* overexpressing *pWOX331*.

that roses might have undergone WGD during long-term evolution, leading to the expansion of the *RcWOX* gene family, enabling them to adapt to various complex growth environments. After analyzing the collinearity within the rose genome, it is found that the number of collinear genes in rose is 0. Therefore, further data evidence is needed to explain the significant expansion of the *WOX* gene family in rose.

4.2 Most of the *WOX* genes had no function during adventitious rooting of rose cuttings

The rooting process of roses is jointly regulated by many genes, but not all members of the *WOX* gene family are involved in this process. Apart from classical clades, most of the *WOX* genes in clades I-V showed no expression during the rose rooting process. Among the 364 *WOX* members from clades I to V, 359 members showed almost no expression during the rooting process of rose, with only *RhWOX276*, *RhWOX51*, *RhWOX33*, *RhWOX284*, and *RhWOX372* genes exhibiting transcriptional expression counts higher than 10 during three or more periods. The classical clades in *RcWOX* gene family members in rose demonstrated similar structures and the presence of UTR in most cases. Conversely, the majority of *RcWOX* family members in clades I-V exhibited UTR loss (Supplementary Figure S1). Similar to classical clades in rose, 14 out of 16 pairs of homologous genes in the soybean *GmWOX* gene family exhibited relatively conserved exon/intron structures (Hao et al., 2019). Many genes in *WOX* gene family of rose did not function during the formation of adventitious roots, while the genes in classical clades exhibited relatively high expression levels, suggesting that these genes might play a role in the rose cutting rooting process. Multiple *WOX* gene family members in different stages of rose rooting responded to cutting signals, such as *RcaWOX1* in *R. canina* callus tissue formation at an early stage (Gao et al., 2014), similar to the expression pattern of *RhWOX185* in *R. 'The Fairy'*. The homologous gene *MdWOX11* of *RhWOX331* in apple cuttings reached its highest expression level at 3 days, and its expression was inhibited by 6-BA (Mao et al., 2023), corresponding with the expression trends of genes *RhWOX372*, *RhWOX316*, and *RhWOX271* in rose. In conclusion, the *WOX* gene of clades I-V regulating the functions of other aspects of roses require further investigation.

4.3 *RhWOX331* in *R. hybrida* can regulate plant meristem activity

Further research on the expression pattern and function of *RhWOX331* in plants revealed that it not only played a role in adventitious root development, but may also be related to plant meristem activity and regulated the development of aboveground and underground parts of plants. Compared to other tissues, the expression level of *RhWOX331* gene in rose roots was significantly increased (Figure 5A), similarly, *WOX* genes in poplar were primarily expressed in roots and leaves (Liu et al., 2014). In *Triticum aestivum*, the homologous gene *TaWOX11* of *RhWOX331* was also highly expressed in roots compared to other

tissues. In addition, both *TaWUS* and *TaWOX9* were transcriptional activators and the transcription activation regions were located at the C-terminus (Li et al., 2020).

Following IBA signaling, the expression of *RhWOX331* was upregulated and its functional role was advanced during the rooting process (Figure 5C). Overexpression of *RhWOX331* in *A. thaliana* demonstrated enhanced primary root and adventitious root formation, indicating the role of *RhWOX331* in promoting primary root elongation and adventitious root development in plants (Figures 6, 7). Similarly, in *A. thaliana*, *AtWOX11* and *AtWOX12* responded to auxin signals, inducing fate transition of stem cells from the pericycle cells to root founder cells, thereby inducing adventitious root formation (Liu and Xu, 2018). *AtWOX11* was involved in the transition of vascular cambium cells to new lateral root primordial cells (Baesso et al., 2018).

The *RhWOX331* promoter, *pWOX331-1* and *pWOX331-2*, triggers GUS protein expression in the meristematic region, indicating the gene's regulation of plant meristematic activity (Figure 8). Additionally, auxin signaling can be detected in this area during *A. thaliana* embryogenesis (Baesso et al., 2018), suggesting that *pWOX331-2* may overlap with auxin signaling to regulate embryonic development. Indeed, during adventitious root formation in *A. thaliana*, the distribution of auxin response coincides with the expression region of *WOX11*, directly responding to the maximum auxin level in the wound-induced pericycle. In rice crown root development, *WOX11* might integrate auxin and cytokinin signaling to regulate the expression of RR2 (Type-A cytokinin-responsive regulator) genes in the crown root primordium, thereby regulating cell proliferation (Zhao et al., 2009). *WOX* gene family played an important role in embryogenesis and shoot apical meristem establishment in conifers (Bueno et al., 2021). Therefore, we propose that *RhWOX331* can respond to auxin signals, regulate plant meristematic activity, and positively correlate with the development of both aboveground and underground parts of plants.

5 Conclusions

The study identified 381 *WOX* genes in *Rosa chinensis* through whole-genome bioinformatics analysis. Phylogenetic analysis and evolutionary tree construction classified the *RcWOX* gene family into eight clades. Gene structure and promoter *cis*-element analysis revealed that genes within the same clade exhibit similar structures and functions. Chromosomal localization of *RcWOX* genes in roses indicated significant expansion on chromosome 2. Relative expression analysis of nine *WOX* gene family members during rose rooting identified several genes with significant expression changes in this process. The *RhWOX331* gene, potentially associated with rooting, was identified through tissue-specific expression analysis, showing high expression in roots and inducibility by IBA while being suppressed by NPA. *RhWOX331* located to the nucleus and exhibited yeast self-activation activity. Overexpression of the *RhWOX331* gene significantly increased the number of lateral roots on the primary root and enhanced the height of *A. thaliana*. Additionally, it accelerated adventitious root formation and alleviated the inhibition of adventitious root

initiation by certain hormones. This gene functioned at the growth point of *A. thaliana* cotyledons. Our study provides initial insights into the role of *RhWOX331* in the process of adventitious root formation in *R. 'The Fairy'*, offering direction and inspiration for future research on the *WOX* gene family of rose.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#), further inquiries can be directed to the corresponding author/s.

Author contributions

LD: Formal analysis, Investigation, Validation, Writing – original draft, Visualization, Writing – review & editing. ZH: Investigation, Validation, Writing – review & editing. WZ: Formal analysis, Writing – review & editing. SL: Investigation, Writing – review & editing. MH: Formal analysis, Writing – review & editing. JZ: Resources, Writing – review & editing. TY: Resources, Writing – review & editing. JD: Methodology, Supervision, Writing – review & editing. DC: Conceptualization, Funding acquisition, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the China Postdoctoral Science Foundation (2023M730535) and National Natural Science Foundation of China (No. 31971700).

References

- Abubakar, A. S., Wu, Y., Chen, F., Zhu, A., Chen, P., Chen, K., et al. (2023). Comprehensive analysis of WUSCEL-related homeobox gene family in *Ramie* (*Boehmeria nivea*) indicates its potential role in adventitious root development. *Biology* 12, 1475. doi: 10.3390/biology12121475
- Baesso, B., Chiatante, D., Terzaghi, M., Zenga, D., Nieminen, K., Mahonen, A. P., et al. (2018). Transcription factors PRE3 and WOX11 are involved in the formation of new lateral roots from secondary growth taproot in *A. thaliana*. *Plant Biol. Stuttg. Ger.* 20, 426–432. doi: 10.1111/plb.12711
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335
- Bannoud, F., and Bellini, C. (2021). Adventitious rooting in populus species: update and perspectives. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.668837
- Bellini, C., Pacurari, D. I., and Perrone, I. (2014). Adventitious roots and lateral roots: similarities and differences. *Annu. Rev. Plant Biol.* 65, 639–666. doi: 10.1146/annurev-arplant-050213-035645
- Bent, A. (2006). *Arabidopsis thaliana* floral dip transformation method. *Methods Mol. Biol. Clifton NJ* 343, 87–103. doi: 10.1385/1-59745-130-4:87
- Bueno, N., Cuesta, C., Centeno, M. L., Ordás, R. J., and Alvarez, J. M. (2021). *In vitro* plant regeneration in conifers: the role of WOX and KNOX gene families. *Genes* 12, 438. doi: 10.3390/genes12030438
- Cao, Y., Han, Y., Meng, D., Li, G., Li, D., Abdullah, M., et al. (2017). Genome-Wide Analysis Suggests the Relaxed Purifying Selection Affect the Evolution of WOX Genes in *Pyrus bretschneideri*, *Prunus persica*, *Prunus mume*, and *Fragaria vesca*. *Front. Genet.* 8. doi: 10.3389/fgene.2017.00078
- Che, G., Gu, R., Zhao, J., Liu, X., Song, X., Zi, H., et al. (2020). Gene regulatory network controlling carpel number variation in cucumber. *Dev. Camb. Engl.* 147, dev184788. doi: 10.1242/dev.184788
- 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
- Cheng, S., Huang, Y., Zhu, N., and Zhao, Y. (2014). The rice WUSCHEL-related homeobox genes are involved in reproductive organ development, hormone signaling and abiotic stress response. *Gene* 549, 266–274. doi: 10.1016/j.gene.2014.08.003
- Cui, W.-H., Du, X.-Y., Zhong, M.-C., Fang, W., Suo, Z.-Q., Wang, D., et al. (2022). Complex and reticulate origin of edible roses (*Rosa*, Rosaceae) in China. *Hortic. Res.* 9, uhab051. doi: 10.1093/hr/uhab051
- De Klerk, G.-J., van der Krieken, W., and De Jong, J. C. (1999). Review the formation of adventitious roots: New concepts, new possibilities. *Vitro Cell. Dev. Biol. - Plant* 35, 189–199. doi: 10.1007/s11627-999-0076-z
- Deyhle, F., Sarkar, A. K., Tucker, E. J., and Laux, T. (2007). WUSCHEL regulates cell differentiation during anther development. *Dev. Biol.* 302, 154–159. doi: 10.1016/j.ydbio.2006.09.013
- Dong, J., Cao, L., Zhang, X., Zhang, W., Yang, T., Zhang, J., et al. (2021). An R2R3-MYB transcription factor *rmMYB108* responds to chilling stress of *rosa multiflora* and

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.2024.1461322/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Conserved motifs and conserved domains of *RcWOXs* in *Rosa chinensis*. (A) Motif composition of *RcWOX* proteins, with different colors representing twenty distinct motifs. (B) Conserved domains of *RcWOXs*, with various colors indicating different structural domains. (C) Green rectangles denote untranslated regions (UTRs); yellow rectangles represent coding sequences (CDS) or exons; black lines indicate introns.

SUPPLEMENTARY FIGURE 2

Cis-acting element analysis of *RcWOXs*. Each *cis*-acting element is indicated by a different color.

SUPPLEMENTARY FIGURE 3

Characterization of *A. thaliana* overexpressing *RhWOX331*.

conferred cold tolerance of arabidopsis. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.696919

Fan, Y., Gao, P., Zhou, T., Pang, S., Zhang, J., Yang, T., et al. (2023). Genome-Wide Identification and Expression Analysis of the Trehalose-6-phosphate Synthase and Trehalose-6-phosphate Phosphatase Gene Families in Rose (*Rosa hybrida* cv 'Carola') under Different Light Conditions. *Plants Basel Switz.* 13, 114. doi: 10.3390/plants13010114

Gao, B., Wen, C., Fan, L., Kou, Y., Ma, N., and Zhao, L. (2014). A *Rosa canina* WUSCHEL-related homeobox gene, RcWOX1, is involved in auxin-induced rhizoid formation. *Plant Mol. Biol.* 86, 671–679. doi: 10.1007/s11103-014-0255-0

Ge, Y., Liu, J., Zeng, M., He, J., Qin, P., Huang, H., et al. (2016). Identification of WOX family genes in *selaginella kraussiana* for studies on stem cells and regeneration in lycophytes. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00093

Groot Crego, C., Hess, J., Yardeni, G., de la Harpe, M., Priemer, C., Beclin, F., et al. (2024). CAM evolution is associated with gene family expansion in an explosive bromeliad radiation. *Plant Cell* koae130. doi: 10.1093/plcell/koae130

Hao, Q., Zhang, L., Yang, Y., Shan, Z., and Zhou, X. (2019). Genome-wide analysis of the WOX gene family and function exploration of gmWOX18 in soybean. *Plants* 8, 215. doi: 10.3390/plants8070215

Hu, B., Jin, J., Guo, A.-Y., Zhang, H., Luo, J., and Gao, G. (2015). GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* 31, 1296–1297. doi: 10.1093/bioinformatics/btu817

Jiang, F. X., Liu, F.-L., and Zhao, L. J. (2012). Overexpression of RaWUS gene of *Rosa canina* regeneration from root tip of transgenic inducing shoot tobacco. *Sci. Silvae Sin.* 47, 43–52. doi: 10.1097/RLU.0b013e3181f49ac7

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., et al. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res.* 47, D1137–D1145. doi: 10.1093/nar/gky1000

Koo, J., Kim, Y., Kim, J., Yeom, M., Lee, I. C., and Nam, H. G. (2007). A GUS/luciferase fusion reporter for plant gene trapping and for assay of promoter activity with luciferin-dependent control of the reporter protein stability. *Plant Cell Physiol.* 48, 1121–1131. doi: 10.1093/pcp/pcm081

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, D1202–D1210. doi: 10.1093/nar/gkr1090

Laux, T., Mayer, K. F., Berger, J., and Jürgens, G. (1996). The WUSCHEL gene is required for shoot and floral meristem integrity in *Arabidopsis*. *Dev. Camb. Engl.* 122, 87–96. doi: 10.1242/dev.122.1.87

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, R., Gao, P., Yang, T., Dong, J., Chen, Y., Xie, Y., et al. (2023). Genome-wide analysis of the SWEET transporters and their potential role in response to cold stress in *Rosa rugosa*. *Horticulturae* 9, 1212. doi: 10.3390/horticulturae9111212

Li, J., Jia, H., Zhang, J., Liu, B., Hu, J., Wang, L., et al. (2018). Effect of Overexpression of *Populus tomentosa* WUSCHEL-related homeobox 4 (PtoWOX4a) on the Secondary Growth of Poplar. *Linye KexueScientia Silvae Sin.* 54, 52–59. doi: 10.11707/j.1001-7488.20180206

Li, Z., Liu, D., Xia, Y., Li, Z., Jing, D., Du, J., et al. (2020). Identification of the WUSCHEL-related homeobox (WOX) gene family, and interaction and functional analysis of taWOX9 and taWUS in wheat. *Int. J. Mol. Sci.* 21, 1581. doi: 10.3390/ijms21051581

Liu, B., Wang, L., Zhang, J., Li, J., Zheng, H., Chen, J., et al. (2014). WUSCHEL-related Homeobox genes in *Populus tomentosa*: diversified expression patterns and a functional similarity in adventitious root formation. *BMC Genomics* 15, 296. doi: 10.1186/1471-2164-15-296

Liu, W., and Xu, L. (2018). Recruitment of IC-WOX genes in root evolution. *Trends Plant Sci.* 23, 490–496. doi: 10.1016/j.tplants.2018.03.011

Lv, J., Feng, Y., Jiang, L., Zhang, G., Wu, T., Zhang, X., et al. (2023). Genome-wide identification of WOX family members in nine Rosaceae species and a functional

analysis of *MdWOX13-1* in drought resistance. *Plant Sci.* 328, 111564. doi: 10.1016/j.plantsci.2022.111564

Mao, J., Niu, C., Li, K., Fan, L., Liu, Z., Li, S., et al. (2023). Cytokinin-responsive MdTCP17 interacts with MdWOX11 to repress adventitious root primordium formation in apple rootstocks. *Plant Cell* 35, 1202–1221. doi: 10.1093/plcell/koac369

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.* 49, D412–D419. doi: 10.1093/nar/gkaa913

Ohmori, Y., Tanaka, W., Kojima, M., Sakakibara, H., and Hirano, H.-Y. (2013). WUSCHEL-RELATED HOMEBOX4 is involved in meristem maintenance and is negatively regulated by the CLE gene FCP1 in rice. *Plant Cell* 25, 229–241. doi: 10.1105/tpc.112.103432

Palovaara, J., Hallberg, H., Stasolla, C., and Hakman, I. (2010). Comparative expression pattern analysis of WUSCHEL-related homeobox 2 (WOX2) and WOX8/9 in developing seeds and somatic embryos of the gymnosperm *Picea abies*. *New Phytol.* 188, 122–135. doi: 10.1111/j.1469-8137.2010.03336.x

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemaître, A., et al. (2018). The *Rosa* genome provides new insights into the domestication of modern roses. *Nat. Genet.* 50, 772. doi: 10.1038/s41588-018-0110-3

Reiser, L., Modrusan, Z., Margossian, L., Samach, A., Ohad, N., Haughn, G. W., et al. (1995). The BELL1 gene encodes a homeodomain protein involved in pattern formation in the *Arabidopsis* ovule primordium. *Cell* 83, 735–742. doi: 10.1016/0092-8674(95)90186-8

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* 3, 1101–1108. doi: 10.1038/nprot.2008.73

Shuang, W., Yang, Z., Meng-Xuan, R., Ying-Ying, L., and Zhi-Gang, W. (2019). Genome-wide analysis of the WOX family reveals their involvement in stem growth of *populus trichocarpa*. *Bull. Bot. Res.* 39, 568. doi: 10.7525/j.issn.1673-5102.2019.04.011

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* 38, 3022–3027. doi: 10.1093/molbev/msab120

Tanaka, W., Ohmori, Y., Ushijima, T., Matsusaka, H., Matsushita, T., Kumamaru, T., et al. (2015). Axillary meristem formation in rice requires the WUSCHEL ortholog TILLERS ABSENT1. *Plant Cell* 27, 1173–1184. doi: 10.1105/tpc.15.00074

van der Graaff, E., Laux, T., and Rensing, S. A. (2009). The WUS homeobox-containing (WOX) protein family. *Genome Biol.* 10, 248. doi: 10.1186/gb-2009-10-12-248

J. M. Walker (Ed.) (2005). *The Proteomics Protocols Handbook* (Totowa, NJ: Humana Press). doi: 10.1385/1592598900

Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S., et al. (2023). The conserved domain database in 2023. *Nucleic Acids Res.* 51, D384–D388. doi: 10.1093/nar/gkac1096

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., et al. (2017). Evolution of rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Mol. Biol. Evol.* 34, 262–281. doi: 10.1093/molbev/msw242

Yu, Y., Yang, M., Liu, X., Xia, Y., Hu, R., Xia, Q., et al. (2022). Genome-wide analysis of the WOX gene family and the role of EjWUSa in regulating flowering in loquat (*Eriobotrya japonica*). *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.1024515

Zhang, J., Eswaran, G., Alonso-Serra, J., Kucukoglu, M., Xiang, J., Yang, W., et al. (2019). Transcriptional regulatory framework for vascular cambium development in *Arabidopsis* roots. *Nat. Plants* 5, 1033–1042. doi: 10.1038/s41477-019-0522-9

Zhang, Y., Jiao, Y., Jiao, H., Zhao, H., and Zhu, Y.-X. (2017). Two-step functional innovation of the stem-cell factors WUS/WOX5 during plant evolution. *Mol. Biol. Evol.* 34, 640–653. doi: 10.1093/molbev/msw263

Zhang, L., Morales-Briones, D. F., Li, Y., Zhang, G., Zhang, T., Huang, C.-H., et al. (2023). Phylogenomics insights into gene evolution, rapid species diversification, and morphological innovation of the apple tribe (Maleae, Rosaceae). *New Phytol.* 240, 2102–2120. doi: 10.1111/nph.19175

Zhao, Y., Hu, Y., Dai, M., Huang, L., and Zhou, D.-X. (2009). The WUSCHEL-related homeobox gene WOX11 is required to activate shoot-borne crown root development in rice. *Plant Cell* 21, 736–748. doi: 10.1105/tpc.108.061655



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural
Sciences, China

REVIEWED BY

Jianhui Ma,
Henan Normal University, China
Ali Razzaq,
University of Florida, United States

*CORRESPONDENCE

Amjad Hameed
✉ amjad46pk@yahoo.com

[†]These authors have contributed equally to
this work

RECEIVED 22 May 2024

ACCEPTED 30 November 2024

PUBLISHED 18 December 2024

CITATION

Ahmed SR, Asghar MJ, Hameed A, Ghaffar M
and Shahid M (2024) Advancing crop
improvement through GWAS
and beyond in mung bean.
Front. Plant Sci. 15:1436532.
doi: 10.3389/fpls.2024.1436532

COPYRIGHT

© 2024 Ahmed, Asghar, Hameed, Ghaffar and
Shahid. 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.

Advancing crop improvement through GWAS and beyond in mung bean

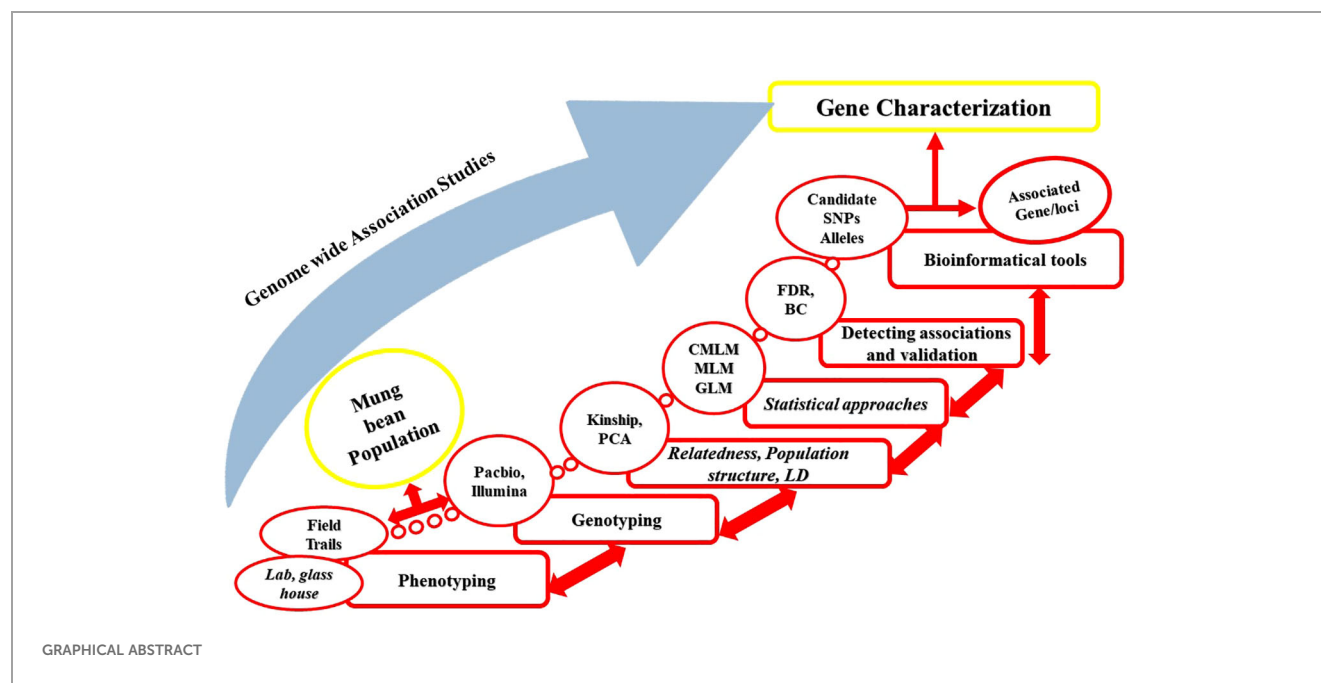
Syed Riaz Ahmed^{1†}, Muhammad Jawad Asghar^{1,2},
Amjad Hameed^{1,3*}, Maria Ghaffar^{1,2†} and Muhammad Shahid^{1,2}

¹Nuclear Institute for Agriculture and Biology College, Pakistan Institute of Engineering and Applied
Science (NIAB-C, PIEAS), Faisalabad, Pakistan, ²Plant Breeding and Genetics Division, Mung Bean and
Lentil Group, Nuclear Institute for Agriculture and Biology, Faisalabad, Pakistan, ³Plant Breeding and
Genetics Division, Marker Assisted Breeding Group, Nuclear Institute for Agriculture and Biology,
Faisalabad, Pakistan

Accessing the underlying genetics of complex traits, especially in small grain
pulses is an important breeding objective for crop improvement. Genome-wide
association studies (GWAS) analyze thousands of genetic variants across several
genomes to identify links with specific traits. This approach has discovered many
strong associations between genes and traits, and the number of associated
variants is expected to continue to increase as GWAS sample sizes increase.
GWAS has a range of applications like understanding the genetic architecture
associated with phenotype, estimating genetic correlation and heritability,
developing genetic maps based on novel identified quantitative trait loci
(QTLs)/genes, and developing hypotheses related to specific traits in the next
generation. So far, several causative alleles have been identified using GWAS
which had not been previously detected using QTL mapping. GWAS has already
been successfully applied in mung bean (*Vigna radiata*) to identify SNPs/alleles
that are used in breeding programs for enhancing yield and improvement against
biotic and abiotic factors. In this review, we summarize the recently used
advanced genetic tools, the concept of GWAS and its improvement in
combination with structural variants, the significance of combining high-
throughput phenotyping and genome editing with GWAS, and also highlights
the genetic discoveries made with GWAS. Overall, this review explains the
significance of GWAS with other advanced tools in the future, concluding with
an overview of the current and future applications of GWAS with
some recommendations.

KEYWORDS

QTLs, mung bean, GWAS, high-throughput phenotyping, structural variants



1 Introduction

Mung bean (*Vigna radiata* L.) is an important food and cash crop in the rice-wheat-based farming systems of Southeast and South Asia and is also cultivated in other regions of the world, especially in the warm regions of the United States, Canada, Australia, and dry parts of southern Europe. Mung bean is native to the Indo-Burma region of Asia, probably first domesticated there, and is believed to have originated in the subcontinent gene center. The wild ancestors of mung bean, *V. radiata* var. *sublobata*, are also from India and can be found in the sub-Himalayan tract, in the Tarai region and in various parts of eastern and western India. Subcontinent is the main center of mung bean diversity, which spreads across the continent from the Himalayas in the north to the southern peninsula and northeastern regions (Mishra et al., 2022). The Indo-Gangetic plains are considered a secondary center of diversity for mung bean. In the past, mung bean seeds were taken by traders and emigrants from Asia to the parts of South America, Latin America, East Africa, Middle East, and Australia (Manjunatha et al., 2023). The area under mung bean cultivation is increasing worldwide and the reasons behind this are its tolerance to heat and drought stresses, low input requirements, high nutritious profile, and most importantly the short crop duration (70 days). Therefore, mung bean has become the most popular niche crop to fill the time gap between wheat (after harvesting) and rice (before sowing). Mung beans thrive in the humid and hot climates of tropical and subtropical regions. They need an annual rainfall of 600 to 870 mm. The best temperature for mung bean growth and development is between 28 and 30°C, though it can tolerate temperatures up to 45°C. The crop is susceptible to waterlogging but can handle slightly salty soils. Mung beans grow well in well-drained loamy to sandy loamy soils with a pH range of 5 to 8 (Sosiawan et al., 2021). Currently, it is cultivated in over six million hectares (6m ha) worldwide which is about 8.5% of the global pulse area and therefore has become one of

the most important edible legume crops (Hou et al., 2019). However, the yield of mung bean in some countries is still very low, ranging from 0.5 to 1.5 t/ha (Hou et al., 2019).

Mung bean is being consumed throughout the world in different forms. The seeds of mung bean are rich sources of proteins, minerals (such as potassium, magnesium and iron), vitamins and dietary fiber compared to other legumes. On dry weight basis the seed of mung bean comprised of 62 to 65% carbohydrates, 3.5 to 4.5% fiber, 4.5 to 5.5% ash, 1 to 1.5% oil and 24 to 28% proteins (Azmah et al., 2023). The proteins of mung bean comprise all the essential amino acids such as lysine, arginine, methionine, tryptophan, isoleucine, valine, phenylalanine, and leucine (Zhang et al., 2024). During sprouting, it has been observed that the proteolytic cleavage of vitamins, amino acids, minerals, and proteins is significantly high. Mung bean holds significant importance in vegetarian diets due to its large and easily digestible proteins. Therefore, mung bean consumption along with other cereals is increasing in the daily human diet (Sehrawat et al., 2024). Mung bean regular consumption not only helps in managing body weight but also provides antioxidant properties, improves digestion, and reduces cholesterol levels in the body to reduce or prevent the risk of chronic diseases. Besides, its nutritious profile, mung bean also plays a significant role in improving soil structure and fertility through nitrogen fixation (Ahmed et al., 2023).

Due to its agronomic and economic importance, it has been used as a model crop to study genomic and genetics studies in other crops of the *Vigna* group. Mung bean is a diploid (2n) in nature with 22 chromosomes and a small genome of around 579 Mb (Somta et al., 2022). In the last few years, research for mung bean has widely expanded since its full genome was sequenced by (Kang et al., 2014). However, its genome has not yet been explored in the ways other models and agronomic crops like *Arabidopsis thaliana*,

rice, wheat, cotton, and maize have been explored. Since mung bean has about 14,187 accessions in the central genebank (the second largest collection in genebank after soybean), it provides an excellent resource to efficiently exploit genetic resources in improving future breeding programs (Schreinemachers et al., 2014). Comparing the re-sequenced genes with the reference genome to check the genetic variations and molecular basis can help in understanding mung bean adaptation to different biotic and abiotic stresses. Moreover, unlike other crop species, the cross compatibility among *Vigna* species has not been widely explored or understood, and so their gene pool. However, there is generally no barrier to cross-compatibility between domesticated cultivars and their close relatives. Some studies have explored wide hybridization to expand the genetic base of *Vigna radiata* using *V. trilobata*, *Vigna umbellata*, and *Vigna mungo*, showing that interspecific barriers can be easily overcome (Lin et al., 2023). Few studies have classified the gene pool of mung bean GP-1, GP-2 and GP-3. The GP-1 consist of *Vigna radiata* and *Vigna sublobata*. The GP-2 consist of *Vigna mungo*, *Vigna umbellata*, *Vigna trinervia*, *Vigna tenuicaulis*, *Vigna stipulacea*, *Vigna grandiflora* and *Vigna subramaniana*. The GP-3 consist of *Vigna angularis* and *Vigna aconitifolia*. Crop improvement has always been the priority of plant breeders (Gayacharan et al., 2020). Crop betterment mainly depends on the availability of genetic variability, which can be found naturally (wild relatives) or induced artificially through hybridization or mutagen. Phenotypic variations within plant species including mung bean are due to the spontaneous natural genetic mutations that are maintained in nature by natural selection, artificial and evolutionary processes. Natural variations have brought great advances in understanding plant physiology, morphology, and its response to adverse climatic conditions. The importance of genetic variation in crop can be understood by elucidating the genetic modifications in agronomic and yield-related traits. For example, pod shattering in mung bean (one of the major issues causing substantial yield loss) is controlled by two quantitative trait loci (QTL) regions (LG1 and LG7). LG7 has also been reported in azuki bean but LG1 is specific in mung bean. Pod shattering in mung bean has been improved through domestication by inducing genetic variation which increased grain yield. Vairam et al. (2017) also reported the improvement of pod shattering in two mung bean genotypes (NM 65 and CO-Gg-7) through induced mutation (Ethyl methane sulphonate and gamma rays) in M2 and M3 generations (Vairam et al., 2017). Genebanks provide a wide source of genetic variation which has been widely used in improving plant species via introducing desired alleles for enhancing yield and developing resistance against biotic and abiotic stresses. On the other hand, modern breeding techniques and domestication processes have also resulted in narrowing down the genetic variation in cultivars that limit crop yield and adaptation.

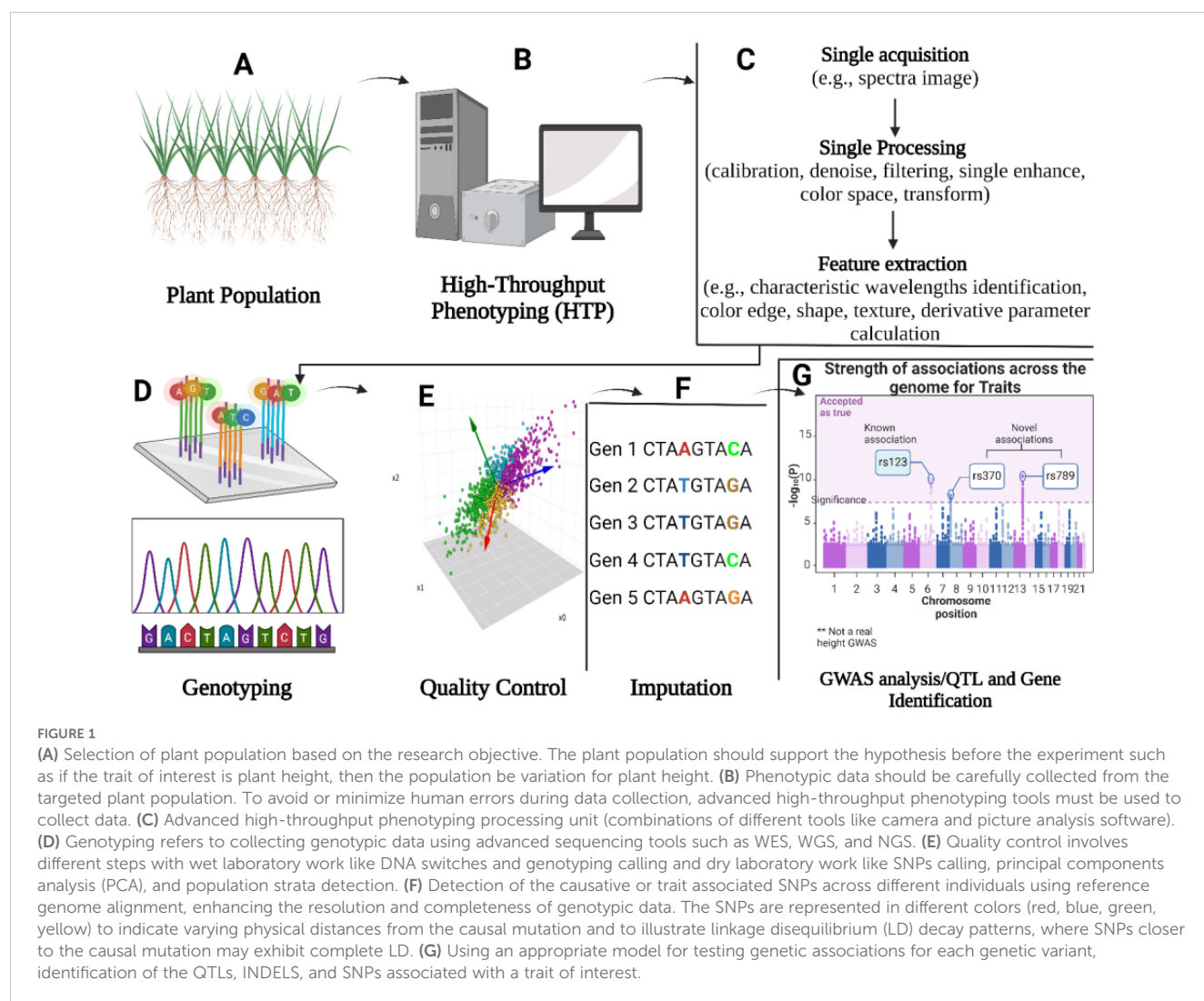
The last two decades have witnessed tremendous computational and technological advances in nucleic acid sequencing. These advances in the field of genome sequencing are due to the simultaneous sequencing of multiple DNA molecules at a high-speed rate and low sequencing cost (Mardis, 2017). Recently, Miga et al. (2020) for the first time presented the gapless telomere-to-telomere fully sequence assembly of the human X chromosome;

before this, thousands of unresolved gaps persisted and no single chromosome was sequenced end to end in any organism. Now, these advances in sequencing technologies have made the genetic improvement of significant traits in mung beans (e.g., early maturity, resistance to mung bean yellow mosaic virus, pod shattering, and seed size) possible. High-throughput-sequencing (HTS) or Next-generation-sequencing (NGS) techniques like genotyping-by-sequencing (GBS) offer the possibility to study thousands of single nucleotide polymorphisms (SNPs) that are associated with the important traits of mung beans. Besides advances in sequencing technologies, numerous excellent statistical-based genetic methods such as whole-genome sequencing (WGS), whole-exome-sequencing (WES) and Genome-wide-association-studies (GWAS) have been proposed to identify genes or alleles controlling target traits. GWAS is a useful technique that can successfully identify the genes of interest for many traits in mung beans as it is based on phenotype and genotype association. In this review we discuss in detail the advancements in GWAS overcoming its limitations, the current status of GWAS in mung bean, discoveries of *k-mers* and structural variations (SVs) as new markers, the status concerning integrating GWAS and high throughput phenotyping in plants (a step forward in unlocking other levels of molecular breeding), expounding the loci found through the multi-scale plant traits obtained by different high-throughput phenotyping techniques in GWAS. In our review, we have focused on mung bean studies as an excellent example of a model pulse crop that has significant genetic improvement due to the identification/discovery of useful novel genes and QTLs, used as markers during selection processes with GWAS. The inherent challenges and future directions are also discussed to enhance our understanding of GWAS, PWAS, and HTP with some guidance for future research.

2 Genome-wide association studies

GWAS detects hundreds of thousands to millions of genetic variants (single nucleotide polymorphism-SNPs) across the genomes of many individuals to identify significant associations between phenotype and genotype. GWAS has revolutionized the field of genetics, especially dealing with complex traits over the past decade. GWAS greatly facilitates analyzing the genetic architectures associated with complex traits and thoroughly explores the genetic basis of phenotypic diversity.

Unlike GWAS in humans, GWAS in plants uses a permanent resource, a population of diverse genotypes that can be re-phenotyped for several traits and only needs to be genotyped once and one can subsequently generate specific mapping populations for particular traits or QTLs (Huang and Han, 2014). The basic theme of GWAS is to compute the association between markers and phenotypes of interest from a diverse panel. The effectiveness and robustness of GWAS in dissecting quantitative traits in crops including mung bean has been fully demonstrated and, is expected to be more effective in identifying the causative gene/loci(s) for complex traits by utilizing recently available large population and high-throughput sequencing technologies. A large



number of alleles (detected through GWAS) and historical recombination events can be used to generate a high-resolution genetic map (Rafalski, 2010) (Figure 1). In association mapping populations, historical-recombination events that assembled through several generations with the help of historical Linkage Disequilibrium (LD) which persist among the representative accessions and enhance association analysis resolution via rapid LD decay (Jaiswal et al., 2019).

GWAS maps quantitative traits and dissect natural genetic variation in combination with genotyping platforms in different crops including mung bean. For example, In GWAS analysis, the use of gene-based 9k SNPs Illumina™ chip provides a higher-genetic resolution that helps in identifying new alleles that improve crop quality, adaptation, and productivity (Thabet et al., 2021) (Figure 2). In mung bean, GWAS will be more informative and robust if we use the newly generated 50k Illumina Infinium iSelect genotyping array. The primary objective of conducting GWAS is to identify causal factors for a given trait and determine the genetic architecture of a specific trait. Crop traits can have either simple genetic architecture (controlled by a low number of loci e.g., mung

bean seed color) or complex genetic architecture (controlled by a large number of loci e.g., mung bean lobed leaflets).

Several steps have been taken so far to improve GWAS methodology but some factors still exist that limit the power of GWAS.

2.1 Factors limiting GWAS power

Many factors limit GWAS's power to detect true associations between phenotype and genotype. Some of the factors are described below:

2.1.1 Variation in phenotypic data

The raw phenotypic data should be carefully analyzed with outliers identified before performing GWAS. The high level of variation in the data from normal variation data points can limit the power of GWAS and might result in false positive or false negative associations. If there are outliers in the phenotype data, the next step should be to assess the impact of these outliers on the

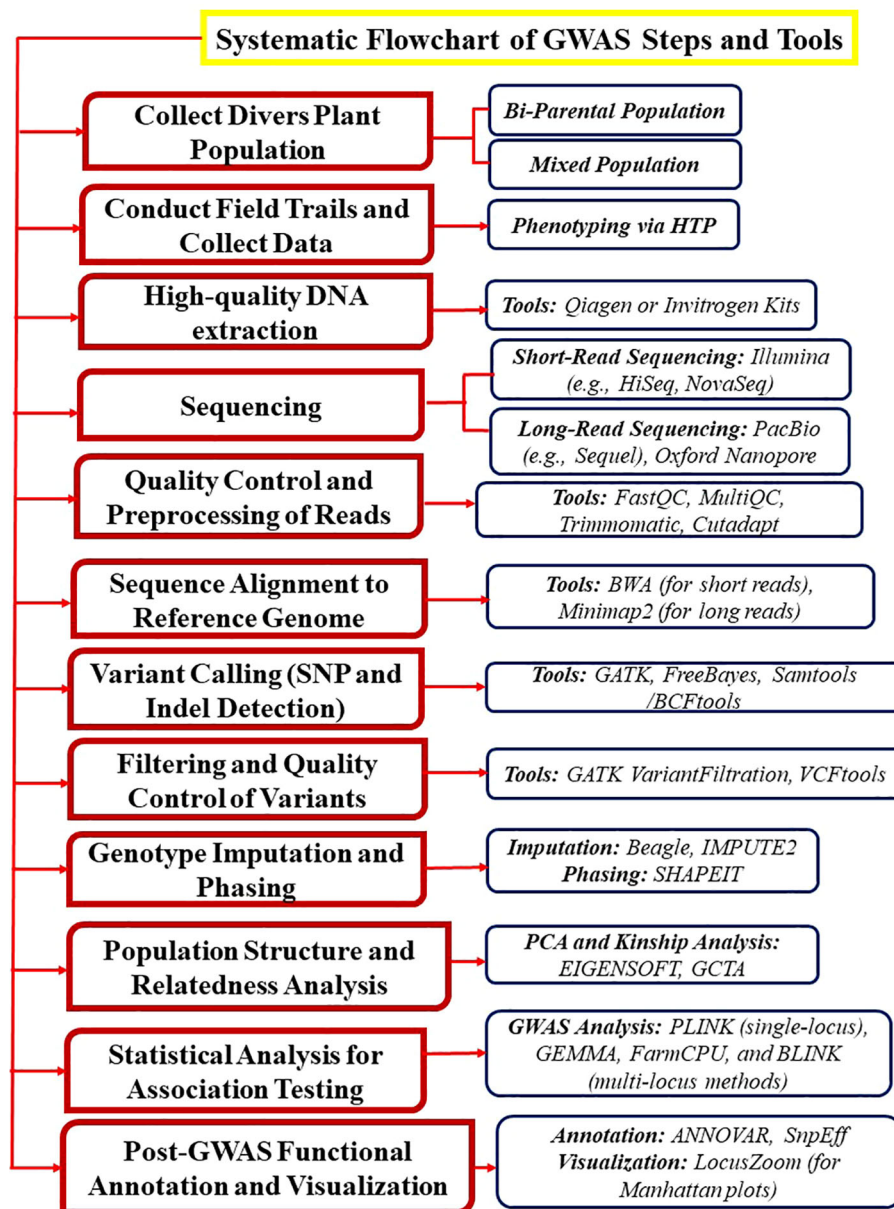


FIGURE 2

This illustration explains the steps and tools involved in performing GWAS in mung bean and other crops. The process begins with the collection of a genetically diverse plant population (e.g., bi-parental or mixed populations). Next, field trials are conducted, and phenotypic data for traits of interest is collected using high-throughput phenotyping (HTP) techniques. High-quality DNA is then extracted using Invitrogen kits, followed by sequencing with advanced platforms such as PacBio. Finally, various analytical tools are applied to identify the associated SNPs.

GWAS. The boxplot is used to test the effect of outliers and visualize the data and if there are extreme outliers in the data they should be excluded. While performing all these steps and removing outliers, it should be highlighted that the removal of outliers should not affect the phenotypic variance as it is very important for association. Additionally, once the filtration of data is completed, traits with high or moderate heritability must be considered for GWAS because heritability is one of the great indicators of how strong the phenotype is associated with genotype and how much the genetic variance has been contributed to phenotype. The power of GWAS to detect true associations among phenotype and genotype is also affected by low broad-sense heritability.

2.1.2 Total number of individuals in the whole population

Population size or sample size is considered a key factor while performing GWAS as obtaining meaningful results is completely dependent on the sample size. Population size is important for explaining portions of genotypic and phenotypic variance; therefore, an increase in sample size will enhance the chances of having true associations, overcoming rare variants, and an acceptable frequency within the population. Sample size ranging from 100 to 500 (or > 500) individuals is needed or acceptable for performing GWAS and the sample size below 100 is considered as a disadvantage that reduces or limits the power of GWAS. Selection

of the individuals from a large population for GWAS may be based on the researcher or breeder's trait of interest, genetic background, growth habit, biological status, and geographic region or location. Mostly, the variation among the individuals within a population can be accessed through phenotypic observation but genotypic information can also be used to access the genetic variation. If the extensive genetic information of individuals is not available, even though their genetic diversity can be estimated through genetic markers (DNA markers) for some of the important traits such as plant height, clusters per plant, pods per cluster, early and late maturity and photoperiod response like in case of mung bean. Once the genotypic and phenotypic analyses of individuals are completed, the individuals with maximum variation are selected for the study. This careful selection of the individuals from the population can detect novel true associations due to greater genetic variation that can be utilized in different aspects of future breeding programs.

2.1.3 Population structure

Population structure is one of the most important components of GWAS. It is a statistical approach/method that calculates or infers the relationship between individuals within a population. It is essential to consider the genealogical or historical relationship between individuals as it affects the analysis and interpretation of results. Since not all individuals are equally related to one another at a genetic level, this is considered the major limitation of GWAS. If the population structure is ignored during performing GWAS or not corrected, it results in spurious associations between the phenotype and genotype. STRUCTURE, a computational-based software (freely available, latest version V. 2.3.4) that is used to describe or address the population structure by generating clusters (subpopulation) within a population (also called Q-matrix) to estimate which individual belongs to which subpopulation. STRUCTURE uses multi-loci data of the genotypes and generates highly accurate clusters to describe population structure. Controlling population structure is always the biggest challenge to be tackled properly. Most of the time, the structured associations are removed to control population structure because of limitations in explaining the total number of clusters and assigning each individual to each cluster but that is not always the adequate way. Moreover, structure analyses are always time-consuming and require rigorous computational analysis. Price et al. (2006) introduced another statistical method (called EIGENSTRAT) for addressing or controlling population structure through principal component analysis (PCA) by reducing the dimensional genotype data (Price et al., 2006). The EIGENSTRAT approach uses genotypic data to estimate genetic variations which are described via a small number of dimensions. Yu et al. (2006) introduced the mixed-model technique for controlling spurious associations by considering multiple/several levels of relatedness through a pairwise relatedness matrix (also known as the Kinship matrix denoted by K) (Yu et al., 2006). The kinship matrix uses the genetic information of individuals to calculate or estimate the relationship or relatedness between a pair of individuals. If the value for the relationship between the individuals is high, it means

that there is a high genetic similarity between these individuals. For example, individuals from the same geographical regions will have the same level of tendency and therefore be clustered in a similar group. The majority of studies conducted so far in mung bean and other crops have used both PCA and STRUCTURE approaches to validate their results (Sokolikova et al., 2020; Wu et al., 2020a; Reddy et al., 2021; Abou-Khater et al., 2022). Sometimes ADMIXTURE software is also used. PCA represents results in a scatter plot by estimating the total variation among the individuals based on their genetic information. If genotypes are randomly distributed within a plot and generate no group, it means that the population has no population structure. STRUCTURE software plots subpopulation against delta k to determine the population structure. STRUCTURE HARVEST is an online website that is used to compress and upload the output results file of STRUCTURE. This software not only provides the acquired population information but also the best k for the proposed population. Table 1 outlines the list of software used in GWAS. Below is the link to STRUCTURE HARVEST

(<https://taylor0.biology.ucla.edu/structureHarvester/>).

2.1.4 Distribution of allelic frequency

Another important component that limits GWAS power is the distribution of allelic frequency; as only a few alleles/loci are present in a few individuals against the whole population. If the number of alleles is fewer or rare, it results in low-resolution power. Thus, allele frequency analysis and distribution directly affect the phenotypic and genotypic associations. If functional alleles are present in the population with low frequency, their detection becomes very challenging unless they have a major effect on the phenotype. If one ignores allelic frequency during GWAS, this might lead to false results. The majority of studies in GWAS focus entirely on common/rare variants and mostly display the allelic frequency at >5%. It means that if the entire population comprises 500 individuals, only 25 individuals are carrying that allele. It shows that this variant is rare with minor allele frequency (MAF) at <5%. This MAF or rare allele explains the variation only in a particular group of individuals within the entire population however, this variant/allele could be important and helpful in future breeding programs. For instance, Youssef et al. (2017) studied a barley population comprised of 209 accessions out of which 13 accessions were collected from East Asia (Youssef et al., 2017). They reported that the 11 accessions from East Asia (out of 13) were carrying the allele (MAF <5%) that significantly affected several complex traits like greater leaf area, number of leaves, and number of tillers. This finding indicates that low-frequency alleles/loci can have immense effects on complex traits. They also proposed that population structure must be carefully studied and linked with GWAS outputs to interpret the results. However, the lower MAF also impacts the ability to detect and utilize the genetic variants associated with the trait of interest. Low MAF also reduces the statistical power to identify the significant association between the traits and alleles. Low MAF increases the chances of false negative results during SNPs association with the trait of interest and thus the reliability of the results gets reduced.

TABLE 1 List of recently developed efficient software for GWAS and genetic analysis.

Software/programs/tools	Application/use	Reference
SMR	Figure out whether the trait and SNP associations are mediated by gene expression levels using Mendelian randomization approach	
Mendelian randomization	Evaluation of causal relation among traits based on genetic overlap utilizing statistics summary of GWAS as input file	(Burgess et al., 2015)
PLINK/PLINK2	Use in different steps while performing GWAS, especially in quality control such as filtering SNPs to separate the associated SNPs from bad SNPs using Hardy Weinberg equation, minor allelic frequency, and genotyping call rate.	(Purcell et al., 2007)
MACH/Minimac	Use to impute missing genotypes adjacent to an available reference panel matched for ancestry and Minimac involved in speeding imputation time.	(Scott et al., 2007)
BEAGLE	Use to impute missing genotypes adjacent to an available reference panel matched for ancestry	(Browning et al., 2018)
GATK	Use for selection of indels and SNPs; acquire reference genome as input file	(Liu et al., 2022b)
IMPUTE ₂	Use to impute missing genotypes adjacent to an available reference panel matched for ancestry; implement more memory when compared to other tools used for imputation	(Howie et al., 2011)
RICOPILI	Use for quality control of raw genetic data and in meta-analysis it requires statistics summary as input file	(Lam et al., 2020)
PLINK	Use to filter the SNPs to minimize the chances of error and identify the real associated SNPs, mostly used after using GATK for further filtration of SNPs	(Han et al., 2022)
BWA-MEM	Use to map reads to the assembled sequence	(Liu et al., 2022a)
SMART-PCA	Use for raw genotypic/sequencing data PCA; provides PCA at the individual level that helps in correcting population stratification	(Kinnersley et al., 2015)
Hisat2	Use to read mapped clean reads to reference genome file	(Liu et al., 2022b)
FastGWA	Used for mixed model genetic association analysis	(Jiang et al., 2019)
BGENIE	Use for continuous phenotypes genetic association: analyses extremely large sample size than is > 100,000; custom made for UK Biobank BGENv1.2 file format	(Bycroft et al., 2018)
SNPTTEST	Use for testing SNPs or genetics associations, perform well with IMPUTE ₂	(Band and Marchini, 2018)
Softonic	Use for statistical data analysis and mostly for principal component analysis (https://origin-1.en.softonic.com/)	(Liu et al., 2022b)
FlashPCA	Similar to SMART-PCA but faster and more scalable with increasing sample sizes compared to SMART-PCA	(Abraham et al., 2017)
PowerMarker		
BamTools/FreeBayes variant caller	Use to call SNPs from raw sequencing or fine genotyping data using reference genome panel (https://github.com/ekg/freebayes)	(Rajendran et al., 2021)
PrediXcan	Using GWAS statistical summary as input file to Prioritize likely causal genes based on transcription data	(Gamazon et al., 2015)
STRUCTURE	Use for structure analysis in GWAS population	(Han et al., 2022)
KMC	Use to estimate the distribution of K-mers across the genome with different parameters	(Liu et al., 2022a)
GenomeScope	Use to estimate genome size, acquire GWAS raw sequencing file as input	(Liu et al., 2022a)
REGENIE	Use for analyzing a large population (>100,000) genetic association and has the ability to assess multiple phenotypes at once; memory effective and rapid	(Mbatchou et al., 2021)
QTL Tools	Use for QTLs identification and analysis; required raw genomic sequenced data as input	(Delaneau et al., 2017)
LDSC	Partitioned SNP-based heritability analyses showing enrichment in sets of functionally related SNPs	(Bulik-Sullivan et al., 2015)
DEPICT	Use predicted gene functions to assess enriched pathways and systematic prioritization of genes	(Pers et al., 2015)

(Continued)

TABLE 1 Continued

Software/programs/tools	Application/use	Reference
Power Marker/SNPhylo	Uses SNP data to develop an un-rooted phylogenetic tree	(Reddy et al., 2021; Sandhu and Singh, 2021)
MAGMA	Use regression framework with competitive testing to assess gene-set and gene-based analysis; permits custom gene sets testing including s options for conditional and interaction testing between gene sets	(De Leeuw et al., 2015)
LDPred-2/LD Pred/PRSCs/SBayesR	Estimation of posterior effect sizes of SNPs using a Bayesian shrinkage approach	(Vilhjálmsdóttir et al., 2015; Privé et al., 2020)
VCFtools	Use to identify chromosomal regions possessing high genetic differences or maximum nucleotide diversity among subpopulations	(Han et al., 2022)
GenomicSEM	Use to assess multivariate genetic correlation using GWAS-based summary statistic	(Grotzinger et al., 2019)
LAVA	Use to assess local multivariate genetic correlation using GWAS-based summary statistic	(Werme et al., 2021)
p-HESS	Use to assess local SNP-based heredity and genetic correlation using GWAS-based summary statistic	(Shi et al., 2017)
superGNOVA	Use to assess local genetic correlation using GWAS-based summary statistic	(Zhang et al., 2020)
fastPHASE	Use to detect SNP markers with MAF 0.05 (http://stephenslab.uchicago.edu/software.html)	(García-Fernández et al., 2021)
SumHer	Use to assess genetic correlation between phenotypes using summary statistic as input; possess several other functions too including assessment of selection bias and partitioned SNP-based heritability	(Speed and Balding, 2019)
GCTA	Use to assess the genetic correlation between phenotypes using raw sequencing file as input	(Yang et al., 2011)
BLUP	Use for different tasks in GWAS such as statistical analysis, association mapping, etc.	(Sandhu and Singh, 2021; Abou-Khater et al., 2022)
FUMA	Use for functional annotation of transcriptomics, proteomics, genomics, and also regulatory regions such as chromatin interaction information and integrates and visualizes all output	(Watanabe et al., 2017)
ANNOVAR and VEP	Use for functional annotation of transcriptomics, proteomics, genomics, and also regulatory regions	(Mclaren et al., 2016)
HaplotypeCaller	Use to identify potential variants in individual samples and generate results in the GVCF file	(Han et al., 2022)
METAL	Use GWAS statistics summary file as input for weighted meta-analysis	(Willer et al., 2010)
GWAMA	Use for Fixed and random effects meta-analysis; allows the specification of different genetic models	(Mägi and Morris, 2010)
FINEMAP	Use to calculate effect sizes and heritability owing to likely causal SNPs; draw statistical-fine mapping acquiring GWAS summary statistics as input file	(Benner et al., 2016)
SuSIE	Use GWAS statistical summary for fine mapping and LD information from a reference panel; based on a Bayesian modification of a forward selection model	(Wallace, 2021)
PAINTOR	Use GWAS statistical summary for fine mapping and functional genomics data for prioritizing likely causal variants	(Kichaev et al., 2014)
GAPIT	Use to perform statistical analysis such as PCA and also develop genetic kinship matrix performing GWAS	(Gela et al., 2021)

2.1.5 Linkage disequilibrium

In a given population if the alleles are associated non-randomly, this is called linkage disequilibrium. LD is another important factor that needs to be considered carefully during GWAS analysis, particularly when defining intervals of tightly associated SNPs which help in explaining the foremost significant loci. If one ignores the alleles' non-random association at different loci, then both causative and non-causative alleles will be incorporated during

analysis and will result in false associations. LD is very important in finding all the markers acquired for covering or scanning the whole genome by determining the distance among loci with the help of LD. If the value of LD is high it means that a small number of markers are required to cover the whole genome (Semagn et al., 2010; Mathew et al., 2018). Long-range LD enhances the chances of spurious associations therefore calculating LD at the beginning of association analysis is necessary to avoid false/spurious associations.

The coefficient of LD can help in measuring the values of how likely two loci are associated and share recombination and mutation history. This analysis is performed using a disequilibrium matrix which displays pair-wise calculations between loci by utilizing the two most common statistics D' and r^2 to measure LD (Flint-Garcia et al., 2003). Several LD analyses performed in plants to date have concluded that D' is likely to be affected by MAF and population size while r^2 is a strong value for estimating how QTL of interest and loci are correlated. LD is likely to be used for estimating the association values (D' or r^2 , >0) between loci as it is important to link the causative SNP with phenotypic variation. It is necessary to consider LD within SNPs as well as in causative alleles during statistical analyses because these analyses reveal whether SNPs identified within LD are significantly associated with a phenotype or not. At this stage in such analysis, it is recommended to consider all SNPs above the threshold level (sometimes every single SNP even below the threshold level) to determine which SNP can clearly explain phenotypic variation since not every highly-associated SNP can have a greater impact on phenotype. SNPs within LD having an r^2 value > 0.2 must be considered for statistical analysis because they might be useful to detect causal loci, especially for those QTLs that are present in the centromeric region (Nadeem et al., 2024).

Mapping resolution (i.e., total markers and density of a given population) in GWAS is of great importance and it is identified through genome size and LD-decay (the rate at which LD declines with physical or genetic distance). The rate of LD decay over a distance (physical/genetic) varies dramatically for loci within a population, within a genome, and among species. To accelerate the rate of LD decay, a greater number of markers would be required for whole-genome association analysis. This LD decay rate helps find the total number of markers required for GWAS by dividing the genome size by the distance at which LD is decayed (Fedoruk, 2013). LD decay in self-pollinated crops such as mung bean is always larger compared with cross-pollinated crops like maize and therefore requires a few markers to cover the whole genome. In mung bean, the LD decay for cultivated and wild species is estimated at about ~ 100 and ~ 60 , respectively (Noble et al., 2018).

If one is interested in estimating the historical recombination events within a particular species then LD pattern analyses within a population can help. However, this depends on several factors like population structure, population size, genotype selection, genetic drift, mutation rate, random mating, recombination rate, and allele frequency. In an association panel (i.e., in artificial selection by researchers), the allelic frequency is not expected to fit with the Hardy-Weinberg principle (HWP) proportion for a given loci (i.e., unlike bi-parental population, genotype frequencies cannot be predicted by association population allele frequencies). However, SNPs that do not fit in HWP are usually excluded from GWAS analysis (Anderson et al., 2019b). In cross-pollinated species, LD decay occurs more rapidly than in self-pollinated species because of large effective recombination. Recombination events in association populations gathered over generations enhance mapping resolution due to a greater number of alleles. If the population size is small, there is a possibility that genetic drift may result in the loss of rare alleles as well as an increase in LD levels. In addition, selection can also increase the level of LD such that if recombination or mutation

occurs among neighboring alleles, they will both be under selection pressure. Thus, association population selection can result in alleles that control specific phenotypes (locus-specific linked alleles) which usually appear in LD. Moreover, migration also increases the level of LD in the population and greatly affects the genetic structure of the association panel. Ignoring genetic drift, migration, mutation, and selection could lead to alleles in linkage equilibrium (D' or $r^2 = 0$). Therefore, critical estimation of population structure and identification of subgroups at the beginning of analyses can reduce all these factors.

2.2 Newly introduced approaches for improving and enhancing GWAS power

The introduction and improvements of new approaches for GWAS have always been an area of interest since LD-based association mapping was first presented (Lander and Kruglyak, 1995). So far, three major areas have been highlighted with the notion that these will not only overcome the above-mentioned limitations but also improve GWAS in different aspects. The three evolving areas include; (1) the development of new efficient marker systems (recently discovered *k-mers* and structural variants (SVs) for genotyping with emphasis on the use of pan-genomics, (2) continuous development and improvements of software and statistical models for statistical analysis to enhance GWAS resolution, and (3) to minimize errors from phenotypic data by introducing high through-put phenotyping techniques (Gupta, 2021b). Simple sequence repeats (SSR) were the first type of markers used in GWAS followed by haplotypes and SNPs. SNPs are the most common type of markers used in GWAS these days. Recently two new classes of markers, *k-mers* and SVs including chromosomal rearrangements (inversions/translocations), insertions/deletions (InDels), presence/absence variation (PAV), and copy number-variations (CNVs) are receiving attention from scientists because they are becoming valuable resources for GWAS.

2.2.1 Genome and GWAS to pan-genome and PWAS

Advances in next-generation technologies (NGS) have made it possible to score thousands of SNPs in a single genotype from an accession panel of species and compare the genome sequence of each genotype with an available reference genome. However, this method cannot score the entire genetic variation present in the genomes of all genotypes of an accession panel used for GWAS. To overcome this issue, it was decided to take advantage of the available genome sequences of individuals within a species, assemble pan-genomes, and use them for GWAS. Tettelin et al. (2005) assembled the first pan-genome in *Streptococcus agalactiae* followed by the development of pan-genomes in plants, animals, and humans (Bayer et al., 2020). Now these pan-genomes are being used as novel reference genomes for GWAS, like the recent acronym PWAS (pan-genome wide association studies) has also been used for GWAS (Manuweera et al., 2019). The applications of *k-mers* and SVs based on early pan-genome studies discovered two key

findings; first, in every species there is about 15 to 40% variable gene content, and second, the genes concerned with *k*-mers and SVs are frequently associated with every type of trait including resistance to abiotic and biotic stresses in crops (Gupta, 2021b). Genomic variations within species are found in both gene content (e.g., PAVs of genes, CNVs distribution across genome, and tandem duplicated genes) and repeated genome portions (e.g., centromere repeats, knob repeats, and transposable elements). This variation has been characterized into three components; core fraction (genomic fraction common to all genotypes within a species), dispensable fraction (which might present in the genome of some genotypes but not in all genotypes) and unique fraction (which is unique to an individual genotype within a species). Till now, several pan-genomic studies have been conducted in different crop plants such as barley (Jayakodi et al., 2020; Wu et al., 2022), wheat (Walkowiak et al., 2020), sorghum (Ruperao et al., 2021), rapeseed (Song et al., 2020, Song et al., 2021), soybean (Li et al., 2014), rice (Zhao et al., 2018a), tomato (Gao et al., 2019), *Brassica oleracea* (Golicz et al., 2016; Bayer et al., 2019), *Brachypodium distachyon* (Gordon et al., 2017) and *Arabidopsis thaliana* (Alonso-Blanco et al., 2016; Van De Weyer et al., 2019). However, no study on pan-genomics in mung bean has been reported yet. There is a need for pan-genomics studies in mung bean to explore the complete genetic variations of some interesting traits such as early maturity and seed size for developing early maturity varieties with large seed size.

2.2.2 Characterization of *k*-mers and SVs for GWAS

During the last few years *k*-mers and SVs have been intensively used for GWAS since pan-genomics have witnessed producing millions of *k*-mers and SVs in single plant species. *K*-mer usually refers to a subsequence in any sequence with a certain length. *K*-mers (they can be in billions to trillions within a species) depend on the *k* value. *k* is the number of nucleotides utilized to develop a set of *k*-mers (Figure 3). For example, AGAT is the sequence of four nucleotides present in DNA, so the value of *k* will be $(4)^k$; therefore, if *k* = 2 then the number of possible *k*-mers is 16, if *k* = 3 then *k*-mers are 64 if *k* = 6 then *k*-mers are 4096 and if the *k* value is 15 or 20 then the *k*-mers will be in billions and trillions, respectively. The value of *k* can be between 2 to 35 or maybe more. *k*-mers with different lengths have already been used for GWAS and pan-genome assemblies. *k*-mers are capable of detecting a wide range of polymorphisms without requiring any reference genome and can be used for GWAS. Before *k*-mer utilization in GWAS, deciding on the size of *k*-mers is the first step (Gupta, 2021a). After this, *k*-mers are isolated from short, sequenced reads (acquired from each genotype of an association panel) and then used for *k*-mers genotyping of one or more association panels. *k*-mers genotyping refers to counting each *k*-mer with a particular size (as mentioned above) in each genotype of the association panel. The genotypic and phenotypic data are then used to identify marker-trait associations (MTAs) in the form of *k*-mers just like SNPs. Voichek and Weigel. (2020) expanded the genetic variants detected through GWAS to include major rearrangements, insertions, and deletions (Voichek and Weigel, 2020). They directly used raw sequence data files and derived *k*-mers and short sequences

as these can mark a huge polymorphism without using a reference genome. Later, they linked *k*-mers associated with phenotype to specific genomic regions. Using this technique, they studied 2000 traits in maize, tomato, and *Arabidopsis thaliana*. Results revealed that MTAs detected through *k*-mers were not different from those detected through SNPs, but *k*-mers allowed detection with more statistical power as compared to SNPs. However, some of the MTAs identified through *k*-mers were not detected earlier using GWAS. They also detected some new associations through SVs and missing regions from reference genomes. This study highlighted the importance of *k*-mers and SVs for GWAS by not only improving GWAS power but also detecting associations with more statistical confidence.

Reduction in sequencing cost of both whole genome sequencing and short reads have allowed characterization of SVs (PAV/CNV) more frequently in crops. PAV and CNV detection techniques have been classified into three categories namely split, pair and depth reads (Alkan et al., 2011). The split read technique involves SVs detection within interrupted short read sequences (Alkan et al., 2011). The read pair technique involves the identification of PAV/CNV based on discrepancies in the distance between paired-end sequences relative to their distance in the reference assembly (Alkan et al., 2011). In the read depth technique, against reference genome short reads are mapped, and the relative depth of a sequence at a locus serves as a proxy for copy number in a particular genotype. Initially, hybridization arrays were used to detect variants but with a greater number of limitations. Later, the availability of whole genome sequencing made the detection of variants much easier but still with some minor limitations. However, these shortcomings have already been addressed to some extent.

Recently, a few other techniques have been developed to further improve the PAV/CNV characterization and also leverage the newly developed library preparation techniques, single-molecules maturation, and long-read sequencing. For instance, connecting molecule approaches such as Strand-Seq, Hi-C, and 10x can retrieve long-range information utilizing short-reads via developing linked reads specialized libraries. Single-molecule techniques (Bionano (optical map) and long read sequencings like Oxford Nanopore and PacBio) permit aligning sequences from several individuals and due to different read lengths; missing sequencings in the reference genome can also be characterized. Both of the above-mentioned techniques have allowed the characterization of both intermediate and small-sized SVs (Levy-Sakin et al., 2019). However, SVs greater than 1Mb can be more effectively characterized through optical maps (Levy-Sakin et al., 2019). SVs with millions of copies in each crop species have already been identified and are intensively being utilized for GWAS/PWAS. Wei et al. (2021) presented a comprehensive quantitative-trait nucleotides [(QTNs) including CNV and PAV] map of rice based on eight GWAS cohorts (Wei et al., 2021). They also developed a genome-navigation system (RiceNavi) for breeding route optimization (BRO) and QTN pyramiding and implemented it in the improvement of Huanghuazhan (intensively grown *indica* rice cultivar). Till now, these developments have led to the most comprehensive characterization of PAV/CNV. Ho et al. (2020) have recently provided a comprehensive review of SVs development in the era

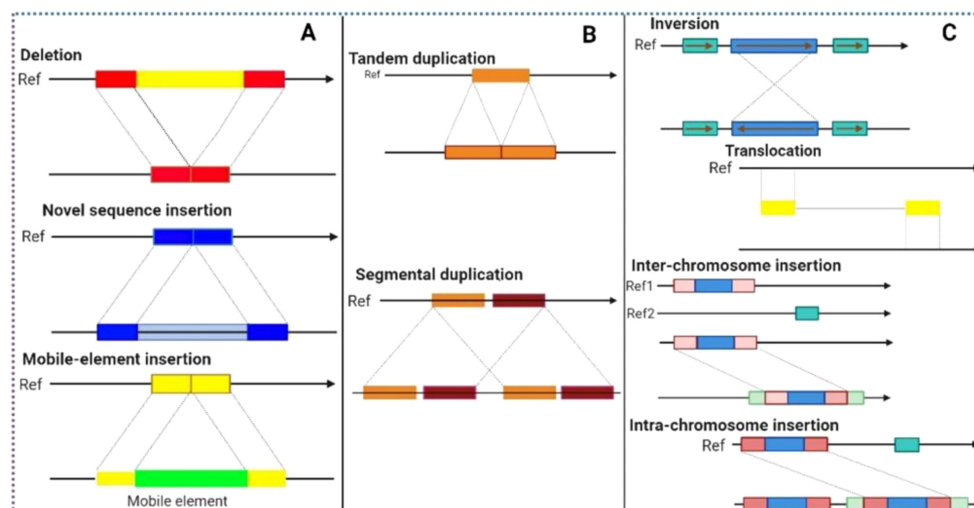


FIGURE 3

Illustration of different structural variants (SVs) that can be found across crop genomes and responsible for creating genetic variations that lead to genetic diversity. Structural variants (such as deletions, insertions, duplications and inversions) in combination with genome-wide association studies (GWAS) can detect hidden SNPs (associated with traits of interest) that remain undiscovered during GWAS analysis.

of genomics (for more information on SVs read the mentioned review).

3 Genetic and molecular advancements in mung bean

Modern genetics, molecular breeding and functional genomics techniques have made plant tolerance against biotic and abiotic stresses easier and faster. Biotic (such as MYMV) and abiotic (drought, salinity and temperature) factors reduce mung bean yield significantly. The emergence and development of the MYMV (through white fly) across India, destroyed the mung bean crop fields completely. Later, this viral disease started spreading rapidly across the borders and started destroying the mung bean crops in other countries like Pakistan and Taiwan. In the early 90s Nuclear Institute for Agriculture and Biology (NIAB), Faisalabad developed the first MYMV-resistant variety through physical mutation (NM-92). The advancement from conventional breeding to mutation breeding (chemical and physical mutagens) was not fast enough as the advancement today in modern genetics techniques. Till now several crops including mung beans have been improved through modern genetics techniques such as marker-assisted breeding, gene silencing, genome editing, QTLs mapping, and NGS. Understanding the crop's genetics associated with the traits of interest allows the molecular breeders to identify the loci and construct a genetic map. Subramaniyan and Narayana (2023), developed a mung bean population through crossing TU 68 (resistant male parent to MYMV) and MDU 1 (susceptible female parent to MYMV), to access the mung bean resistance to MYMV through genetic markers. Some of the introgression lines showed significant resistance to MYMV along with high yield. They further identified the genes associated with the disease resistance through

using genetic markers. Talakayala et al. (2022), employed CRISPR-Cas at two different locations AV1 (coat protein) and AC1 (rep protein) in mung bean to develop resistance against the MYMV. The transformed lines (containing Cas9 cassette) displayed minimal mosaic symptoms and displayed resistance against MYMV by reducing the accumulation of AV1 and AC1. Besides, several studies have identified many genes in mung bean associated with several biotic and abiotic factors and constructed QTL maps. Some of the examples are given below in details. Figure 4 contains the several genes identified associated with traits and their chromosomal location in mung bean.

3.1 QTLs detection in mung bean through GWAS

We have seen the negative effects of climate change on crop growth and development including a significant reduction in yield. Complex traits like yield and seeds per pod in mung bean are controlled by several alleles and therefore it is difficult to understand the underlying genetic architecture of complex traits (Yuan et al., 2020). For example, GWAS analysis in mung bean recently discovered five QTLs associated with resistance to mung bean yellow mosaic virus (MYMV). The QTLs $qMYMV_{10-1}$, $qMYMV_{6-1}$, $qMYMV_{4-1}$, $qMYMV_{5-1}$ and $qMYMV_{4-1}$ was identified on chromosomes 10, 6, 5, and 4 with a total of 538 SNPs covering 1291.7 cM distance. $qMYMV_{4-1}$ (on chromosome 4) was found as major and the most stable QTL for resistance to MYMV (Mathivathana et al., 2019). GWAS analyses have discovered several novel QTLs for various traits and environmental conditions like salinity stress in different crops (including mung bean) that have not been reported previously. Salinity stress is known to cause a major yield reduction in mung

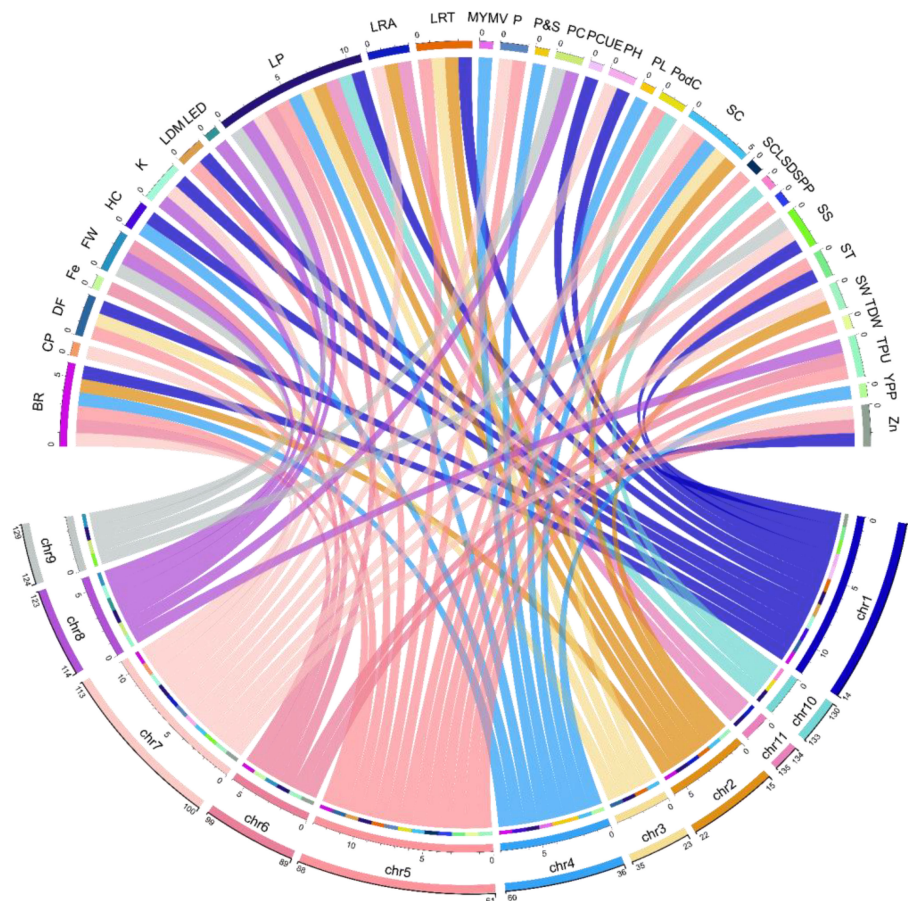


FIGURE 4

Distribution of some of the most important genes across chromosomes associated with different mung bean traits discovered through GWAS. The numbers on each chromosome (in second line) for example on chr1 (1-14), Chr2 (15-22), Chr3 (23-35) represent the number of genes present on the chromosome associated with the above traits (Supplementary Table 1); SC (seed color), BR (bruchid resistance), CP (crude protein), DF (Days to flowering), FW (Fusarium wilt), HC (hypocotyl color), Fe (Iron), LRA (lateral root angle), LDM (leaf drop at maturity), LRT (Leaf related traits), LP (Lectin proteins), LEC (Root length distribution), MYMV (Mung bean yellow mosaic virus), PC (Phosphorus conc.), PCUE (P concentration and P utilization efficiency), P (Phosphorus), PH (Plant height), PC (Pod color), PL (Pod length), K (Potassium), P_S (Quality traits (Protein and starch), SS (Salinity stress), SCL (Seed coat luster), ST (Seed texture), SW (Seed weight), SPP (Seeds per pod), SD (Shoot development), TDW (Total Dry Weight), TPU (Total Phosphorus Uptake), YPP (yield per plant), Zn (Zinc).

bean. Liu et al. (2022a) reported seven QTLs (*EVM0012371*, *EVM0002218*, *EVM0029605*, *EVM0033924*, *EVM0022712*, *EVM0017397*, and *EVM0018329*) significantly associated with salt tolerance in mung bean using GEMMA and EMMAX (Liu et al., 2022b). These QTLs are distributed on chromosomes 1 and 3. They also reported that the expression level of candidate gene *VrFR08* was up-regulated under salinity stress. Furthermore, another study reported 5288 SNPs markers through GWAS to mine alleles associated with salinity stress in mung bean. Significantly associated SNPs and QTLs were identified on chromosomes 7 and 9 with 7 and 30 genes, respectively. However, QTL on chromosome 7 stretched from position 2,696,072 to 2,809,200 bp having seven genes but only one gene *Vradi07g01630* was functionally annotated. Similarly, QTL on chromosome 9 stretched from 19,390,227 to 20,321,817 bp having 30 genes but only two genes *Vradi09g09600* and *Vradi09g09510* were functionally annotated (Breria et al., 2020a). Dissecting the root genotypic and phenotypic variability in mung bean accessions using

GWAS revealed that chromosomes 2, 6, 7, and 11 possess QTLs that control lateral root angel (LRA), chromosomes 3 and 5 having QTLs that control total dry weight (TDW) and volume (VO) and QTLs on chromosome 8 control total root length growth rate (TRLGR). Moreover, gene description on different chromosomes; chromosome 2 has two genes first (–)-Germacrene D synthase-like and second gene description is not given (both genes are significantly associated with LRA), chromosome 3 has one gene Mannose-1-phosphate guanylyltransferase1 (associated with TRLGR), chromosome 5 has one gene dehydration-responsive element-binding protein 2H (DREB2) associated with TDW, chromosome 6 also has one gene associated with LRA but has no description, chromosome 7 has two genes first Beta-galactosidase 3 and second gene description is not given (both associated with LRA). Chromosome 8 possesses two genes first Monodehydroascorbate reductase, second Uncharacterized LOC106771882 associated with LED. Chromosome 11 has one gene Protein FAR1-RELATED SEQUENCE5 associated with LRA

(Chiteri et al., 2022). To this end, several mung bean populations and marker types have been used to study the genetic variability among accessions and a wide range of important traits. For example, the mini core mung bean collection (consisting of 293 to 297 accessions) established by the World Vegetable Centre Taiwan (also called AVRDC) is intensively used for GWAS studies that revealed QTLs for different traits and stress conditions (Breria et al., 2020b; Sokolkova et al., 2020). GWAS output in mung bean provides novel candidate genes and alleles that can be used in future breeding programs to develop resistance to abiotic and biotic stresses and enhance yield to meet targets.

3.2 GWAS: a driver of candidate gene discovery in mung bean

Statistical geneticists commonly believe that GWAS have rendered traditional candidate gene identification techniques obsolete (Duncan et al., 2019). The importance of population association mapping in identifying the candidate genes associated with particular traits can be estimated from the number of studies published since 2015. In this section, we shall also illustrate the potential of GWAS in detecting allelic variations with examples shown in Table 2. The first genome-wide study in mung bean was conducted by Van et al. (2013) to assess the genetic diversity and identify the SNPs markers associated with resistance the MYMV and seed shattering (Van et al., 2013). They used Illumina Hiseq to sequence Gyeonggi jaerae 5 and Sunhwanokdu (two mung bean cultivars) and sequenced more than 40 billion base pairs (from both cultivars) to a depth of 72x. They identified a total of 305,504 SNPs out of which 42 were significantly associated with both the traits mentioned above. In the beginning, identifying candidate genes using whole-genome sequence data was difficult due to the lack of knowledge of GWAS and the tools/software required for handling the large data. Later Korean scientists, Daovongdeuan et al. (2017) carried out the second GWAS attempt in mung bean to study seed size and color using 218 accessions collected from different regions of the world (Daovongdeuan et al., 2017). They could not identify any significant SNP marker associated with the studied traits at a LOD of 6 and p-value <0.05. This second attempt of GWAS in mung bean once again failed in reporting the candidate genes associated with seed size and color. However, they reported that the studied traits were controlled by several alleles but with minor effects. *VrMYB113* (on chromosome 4) and *Vrsf3'h1* (on chromosome 5) are the first two genes in mung bean discovered using GWAS; that are associated with the seed coat color (Noble et al., 2018) (Figure 4, Supplementary Table 1). *MYB113* was first reported by Gonzalez et al. (2008) in *Arabidopsis thaliana*, responsible for anthocyanin biosynthesis (Gonzalez et al., 2008). Anthocyanin concentration in mung bean and other plants depends on the expression levels of *MYB113*. miR828 (micro-RNA828) and TAS4 (trans/acting siRNA4) are small endogenous RNAs, responsible for post-transcriptional suppression of *MYB113*. TAS4 and miR828 mutants were developed using CRISPR-Cas to further confirm the involvement of *MYB113* in anthocyanin biosynthesis. The mutant plants accumulated more anthocyanin compared with

untreated plants, thus confirming the significant association of *MYB113* with seed coat color (Sunitha and Rock, 2020; Koo and Poethig, 2021). *FRO8* gene is another example detected by GWAS, associated with tolerance to salinity stress in mung bean (Liu et al., 2022b). *FRO8* had a direct connection with the *BELL-1* gene. The *BELL-1* like family (*BELL*) of transcription factors is ubiquitous among plant species and found in regulating a range of developmental processes through interacting with *KNOTTED1*-like proteins (Kurt and Filiz, 2020). *Jg5489* which is a homolog of *WUSCHEL*-related homeo-box-3 (*WUS*), associated with yield per plant in mung bean has also been discovered using GWAS. In the same study, they also discovered several other candidate genes *jg35209* and *jg3587* that are homologs to *Glyma09g33350/Glyma09g33340* and *Glyma03g01540* (soybean candidate genes identified using GWAS) associated with days to flowering (Mao et al., 2017).

In contrast, Ahmed et al. (2021), for the first time in chickpeas discovered *RPLP0* and *EMB8-like candidate genes* using GWAS associated with salinity stress (Ahmed et al., 2021). Similarly, Maalouf et al. (2022) discovered candidate genes (*MYB*-related P-like protein, *PsaA*, *RCH1*, *NAK*, and *LRR*) through GWAS in faba beans associated with herbicide tolerance. The successful above-mentioned examples of candidate gene identification through GWAS provide strong evidence that GWAS can rapidly detect hidden loci/genes associated with important plant traits and that can be effectively used to further strengthen the mung bean breeding program.

4 Recent advances in high-throughput phenotyping in GWAS

Domestication started many decades ago in response to feeding the large population and protecting plants from adverse climatic conditions. Domestication requires many years (about 6 to 7 years mostly) to develop a single crop variety. This challenge forced researchers to find new ways to speed up the process of crop improvement. Therefore, various techniques were successfully introduced to improve crops within a short duration and whole genome sequencing was one of those techniques. Since whole-genome sequencing has been achieved in several crops, functional genomics studies have stepped into the big-data and high-throughput phenomics era. In 1911, Wilhelm Johannsen characterized the word phenotype for the first time as “all type of organisms can be distinguished by direct inspection or with finer method of measurements or description” (Johannsen, 1911). Later, Davis in 1949 defined the word phenome as “the total of extra genic, non-auto-reproductive portions of the cell and represented the set of phenotypes” (Davis, 1949). Simply, crop phenomics can be defined as “the multi-disciplinary study of high throughput accurate acquisition and multi-dimensional analysis of phenotypes on a large scale through crop development” (Yang et al., 2020). Plant phenotype is influenced by genotype and environment (G x E) interactions. According to Mendelian genetics, in the presence of a dominant allele, the recessive allele

TABLE 2 List of candidate gene(s) discovered and validated using GWAS in mung bean and other pulses.

Population	Sample Size	Growth habit	Model	Markers	Phenotype	Software/programs/tools	Candidate gene(s)/Gene ID	Chromosome Position	Validation	Reference
Mung bean										
Chinese accessions	112	Summer	GEMMA, EMMA	160.14K	Salinity-stress survival rate 10 and 15 Days	Softonic, Hisat2, and GATK	<i>VrFRO8</i>	Chr.1	Comparative genomics, Transcriptome and Metabolomics PCR, statistical analysis, data integrations	(Liu et al., 2022b)
Chinese and other origin accessions	750	Spring and Summer	GEMMA	2.9K	Insect resistance, yield, gain composition, pod width, pod length, flowering period, etc.	KMC, GenomeScope, BUSCO, BWA-MEM, Hi-C, LTR_retriever, RepeatModeler, RepeatMasker, ADMIXTURE, BRAKER2, HISAT2, ProtHint, GUSHR, Infernal, Barrnap, Rfam, r8s, TimeTree, WGD detector, Profiler, MCScan, MaSuRCA, QUAST, CD-HIT, Mosdepth, Picard, EIGENSOFT, iTOL, R-programming, VCFtools and LDBlockShow	hg22573, hg5284, hg13746, hg35209, hg3587, hg30665 and 250+ others.	Chr.1, Chr.4, Chr.5, Chr.7, Chr.10	Markers, Transcriptome and Metabolomics, statistical analysis, data integrations	(Liu et al., 2022a)
USDA and Asian accessions	375	Growth chamber	MLM	26.5K	TDW, VOL, TRL_GR, LED, LRA, etc.	TASSEL	LOC106755829, LOC106753988, LOC106768494, LOC106776541, LOC106772343, LOC106771882, LOC106772343	Chr. 2, Chr. 7, Chr. 11, Chr. 8, Chr. 5	Mapping, Molecular markers, statistical analysis, data integrations	(Chiteri et al., 2022)
Chinese accessions	558	Spring and Summer	GEMMA	69.9K	Branch number, plant height, pod width, pod length, Flowering time, and quality parameters	SAMtools, GATK, e HaplotypeCaller, PLINK, MEGA-X, STRUCTURE, VCFtools, R-programming	Vradi05g00200, Vradi03g06500, Vradi04g07830, Vradi04g07820, Vradi04g07810, radi04g07800	Chr. 5, Chr. 3, Chr. 4	Re-sequencing, variant Mapping, Molecular markers, statistical analysis, data integrations	(Han et al., 2022)
AVRDC accessions	120	Glass house	MLM, CMLM	55.6K	TDW, PC, TPU, PUtE	TASSELv5.0, STRUCTUREv2.3.4, PLINK,	VRADI01G04370, VRADI05G20860, VRADI06G12490, VRADI08G00070, VRADI08G20910, VRADI09G09030	Chr.1, Chr.5, Chr.6, Chr.8, Chr.9	Sanger sequencing, expression	(Reddy et al., 2021)

(Continued)

TABLE 2 Continued

Population	Sample Size	Growth habit	Model	Markers	Phenotype	Software/programs/tools	Candidate gene(s)/Gene ID	Chromosome Position	Validation	Reference
Mung bean										
						MEGA v6.0, PowerMarker v3.51			analysis, Markers	
USDA accessions	482	Spring and Summer	CLMM, FarmCPU	264.5K	Qualitative seed traits, 100- seed weight, days to flowering, Plant height, etc.	CLML, GAPIT, BLUPs, FarmCPU, Numericware-i, STRUCTURE, PLINK, SNPhylo, DISTRICT, R-programming, CLUMPP, adegenet	LOC106774729, LOC106774729, LOC106758789, LOC106759308, LOC106760769, LOC106764910, LOC106772003, LOC106773047, LOC106774971	Chr.1, Chr.2, Chr.4, Chr.5, Chr.6, Chr.8, Chr.9, Chr.10	Sequencing, Histogram plots, Statistical analysis, Molecular markers	(Sandhu and Singh, 2021)
AVRDC mini-core collection	297	:	MLM, GLM	5.3K	Seed coat luster	TASSEL 5.2.31, STRUCTUREv2.3.4, R-programming	Vradi05g09110, Vradi05g09100, Vradi05g08320	Chr. 5	Molecular Markers, statistical analysis, data integrations	(Breria et al., 2020b)
AVRDC mini-core collection	284	Controlled Conditions	FarmCPU, MLM	5.3K	Salinity stress	TASSEL 5.2.31, STRUCTUREv2.3.4, R-programming	Vradi07g0163, Vradi09g09600, Vradi09g09510	Chr.7, Chr.9	Molecular Markers, statistical analysis, data integrations	(Breria et al., 2020a)
USDA accessions	95	Summer	MLM, GLM	6.48k	Seed minerals Zn, P, S, Mn, K, Fe, Ca	TASSEL, BWA, R-programming	Vradi01g00840, Vradi01g00830, Vradi01g00820, Vradi05g16350, Vradi07g26340, Vradi07g26320, Vradi07g1418, Vradi08g22740, Vradi06g10210, Vradi06g10120, Vradi06g10060, Vradi06g10020, Vradi06g09900, Vradi07g06200, Vradi07g05950, Vradi01g05570, Vradi06g02380	Chr.1, Chr.5, Chr.7, Chr.8, Chr.6	Statistical analysis, Molecular markers, Mapping, Data integrations	(Wu et al., 2020b)
Australian accessions including wild types	482	Summer	MLM	22.2K	Seed coat color	TASSEL, R-programming, DARwin v6.0	VrMYB113, Vrsf3'h1	Chr.4, Chr.5	Mapping, Data integrations, Statistical analysis	(Noble et al., 2018)
Other Species										
Lentil accessions from 60 countries	326	Winter	MLM	164.1K	Aphanomyces root rot index, Root dry weight,	Haploview (v 4.2), Cartographer, BWA, SAMtools, Freebayes (v1.2),	ABCA, PE, and CHI	Chr.2, Chr.4, Chr.5, Chr.7	qRT-PCR, QTLs Mapping,	(Ma et al., 2020)

(Continued)

TABLE 2 Continued

Population	Sample Size	Growth habit	Model	Markers	Phenotype	Software/programs/tools	Candidate gene(s)/Gene ID	Chromosome Position	Validation	Reference
Other Species										
					Shoot dry weight,	VCFtools, BEAGLE (v 3.3.2), R-programming			Molecular markers Statistical analysis,	
ICARDA lentil accessions	176	Winter	GLM	22.5K	Days to first flower, Plant height, Seed per pod, days to maturity, harvest index	TASSEL, Freebayes, BamTools, Stacks, RAD-Tags, PGDSpider, STRUCTURE, UPGMA, NTSYS-PC program 2.02k, CDC Redberry	Marker trait associations (MTAs) SLCCHR3, SLCCHR5, SLCCHR6, SLCCHR7	Chr.2, Chr.3, Chr. 5, Chr. 6, Chr. 7	Molecular markers, PCR, Statistical analysis,	(Rajendran et al., 2021)
Diverse Lentil accessions	200	Winter	MLM	21.6K	Resistance to anthracnose race 1	VCFtools, MSTMap, ICIMapping, KnowPulse database, SNPRelate, STRUCTURE, Bayesian-model-based, GAPIT, R-programming	Lcu.2RBY.3g006340, Lcu.2RBY.3g006380, Lcu.2RBY.3g005880, Lcu.2RBY.3g005310, Lcu.2RBY.3g006350 and MATs includes Lcu.2RBY.Chr6.374326758, Lcu.2RBY.Chr5.437944230, Lcu.2RBY.Chr5.28637458, Lcu.2RBY.Chr4.442702133, Lcu.2RBY.Chr4.442702129	Chr.3, Chr.4, Chr.5, and Chr.6	GenBank, Molecular Markers, QTL mapping, statistical analysis	(Gela et al., 2021)
Lentil accessions	143	Winter	GEMMA	22.2K	Identification of pre-biotic carbohydrates, Total Starch, Resistant Starch, Stachyose +Raffinose, Sucrose, Fructose, Glucose and Mannitol	VCFtools, GAPIT, TASSEL, FarmCPU, VanRaden, PLINK, R-programming	Lcu.2RBY.7g016860, Lcu.2RBY.7g016850, Lcu.2RBY.6g060190, Lcu.2RBY.6g015410, Lcu.2RBY.3g007570, Lcu.2RBY.2g028680, Lcu.2RBY.2g028670, Lcu.2RBY.2g028680, Lcu.2RBY.2g028670, Lcu.2RBY.1g069450, Lcu.2RBY.1g023480, Lcu.2RBY.7g048400, Lcu.2RBY.7g048380, Lcu.2RBY.7g048450, Lcu.2RBY.7g048410, Lcu.2RBY.7g048380, Lcu.2RBY.4g007850,Lcu.2RBY.6g015410, Lcu.2RBY.5g043890, Lcu.2RBY.4g045790, Lcu.2RBY.2g055260,Lcu.2RBY.1g020350, Lcu.2RBY.1g020320, Lcu.2RBY.6g040560, Lcu.2RBY.4g026570, Lcu.2RBY.3g057050	Chr.1, Chr.2, Cr.3, Chr.4, Chr.5, Chr.6, Chr.7	Statistical analysis, Histograms, GBS and molecular markers, and chemical analytical techniques through advanced instruments	(Johnson et al., 2021)
Lentil accessions	118	Winter	MLM	3.2K	Resistance to Pea Aphid (PA resistance traits	STACKS v.2.0, BWA, SAMTOOLS v.0.1.19, BEAGLE v.3.3.2, FarmCPU, HAPLOVIEW v.4.2, R-programming	Marker trait associations (MTAs) including 7173_43, 7453_32, 5957_51, 5443_27, 5421_34, 3884_57, 4584_48, 3782_10, 2642_48, 3385_39	Chr.2, Chr.3, Chr.4, Chr.5	Statistical analysis, Association mapping, PCR, Molecular markers	(Das et al., 2022)
Common bean Spanish diverse panel	308	Spring	MLM	32.8K	Pod morphological	fastPHASE, Tassel, mrMLM, GAPIT,	Phvul.010G118700, Phvul.010G117200, Phvul.008G019500, Phvul.007G206200, Phvul.006G074600, Phvul.002G141800,	Chr.1, Chr.2, Chr.4, Chr.6, Chr.7, Chr.8, Chr.10	Statistical analysis, QTL mapping,	(García-Fernández et al., 2021)

(Continued)

TABLE 2 Continued

Population	Sample Size	Growth habit	Model	Markers	Phenotype	Software/programs/tools	Candidate gene(s)/Gene ID	Chromosome Position	Validation	Reference
Other Species										
					and color characters		Phvul.001G262600, Phvul.001G229900, Phvul.001G139100		Molecular markers, Sequencing	
Faba bean accessions from ICARDA and other countries	140	Winter, Spring	GEMMA	10.8K	Herbicide tolerance traits include Plant height, seeds per plant, pods per plant, branches per plant, yield per plant, days to maturity, and days to 50% flowering	ADMIXTURE, TASSEL, Bowtie, GenStat, BLUP, R-programming	SNP trait association including <i>SNODE_7114_58</i> (gene, MYB-related protein P-like), <i>SNODE_559376_60</i> (gene photosystem I core protein PsA), <i>SNODE_4187_38</i> (gene, malate dehydrogenase), <i>SNODE_3696_16</i> (gene, Probable serine/threonine-protein kinase NAK), <i>SNODE_14298_44</i> (gene, LRR receptor-like serine/threonine-protein kinase RCH1), <i>SCONTIG127798_41</i> (gene, acidic endochitinase)	— —	Molecular markers, Statistical analysis	(Abou-Khater et al., 2022)
Faba bean accessions from ICARDA and other countries	134	Summer, Spring, Winter	GEMMA	10.8K	Heat resistance including, Plant height, seeds per plant, pods per plant, branches per plant, yield per plant, days to maturity, days to flowering, pollen germination, and 100 seed weight	TASSEL, Bowtie, ADMIXTURE, TASSEL, BLUP, R-programming	LOC11440721, LOC113783927, LOC11440721, LOC109335950, LOC11420332, LOC11420332, LOC11420332, LOC11430352, LOC101493666, LOC101512103, LOC114380151, LOC113847809, LOC109813943, LOC25500962, LOC101496898, LOC101496898, LOC112012620, LOC101492966, LOC101492966, LOC11425609, LOC109813943	— — —	Molecular markers, Statistical analysis	(Maalouf et al., 2022)
Faba bean accessions	290	Winter	MLM, GLM	687	Frost resistance traits including AUSPC, LTAF, LCAF, FAC, FPC,	TASSEL, PowerMarker, QTL Network, STRUCTURE, R-programming	Marker trait associations (MTAs) including VF_MT3G086600, VF_MT2G027240, VF_MT4G125100, VF_MT4G127690, VF_MT5G026780, VF_MT4G007030, VF_MT5G005120, VF_MT5G033880, VF_MT7G090890, VF_MT7G084010, VF_MT5G046030	Chr.2, Chr.3, Chr.4, Chr.5, Chr.7	QTL mapping, PCR, Molecular markers, Statistical analysis	(Sallam et al., 2016)
Faba bean a inbred lines	189	Winter	MLM	2.54K	Convicine and vicine contents in seeds	TASSEL, WinISI II, R-programming	SNPs associations, Affx-1003954634, Affx-1003937842, Affx-309473691, Affx-308714105, Affx-309732154, Affx-308989324, Affx-309859410, Affx-309903736, Affx-308750155, Affx-310120776, Affx-309712729, Affx-	Chr.1, Chr.2, Chr.3, Chr.4, Chr.5, Chr.6	Molecular markers, NIR, HPLC, Statistical analysis, Genetic map, QTLs	(Puspitasari et al., 2022)

(Continued)

TABLE 2 Continued

Population	Sample Size	Growth habit	Model	Markers	Phenotype	Software/programs/tools	Candidate gene(s)/Gene ID	Chromosome Position	Validation	Reference
Other Species										
							310628027, Affx-308848038, Vf_Mt4g053880			
Chickpea ICRISAT accessions	280	Winter	MLM	4.6K	Zn and Fe concentrations, Day to 50% flowering, Days to maturity, and 100 seed weight	TASSEL, GAPIT, Admixture, BLINK, STRUCTURE, STRUCTURE HARVESTER, BLUPs, FarmCPU	S6_7891103, S7_9379786, S4_4477846, S6_26554579, S4_31996956, S1_2001361, S1_2772537, S7_32973784	Chr.1, Chr.4, Chr.6, Chr.7	Molecular markers, Statistical analysis, Mapping,	(Srungarapu et al., 2022)
Chickpea Australian accessions	315	Winter	MLM	298K	Yield related traits	BLUE, GeneStat, Tassel, R-programming	Evaluated the genetic variability based on the number of SNPs per chromosome. No gene was reported.	26.7K SNPs on Chr.1, 18.1K on Chr.2, 13.6K on Chr.3, 52.3K on Chr.4, 27.8K on Chr.5, 129.3K on Chr.6, 25K on Chr.7, 5.5K on Chr.8	Statistical analysis,	(Liu et al., 2021)
ICARDA Chickpea accessions	186	Winter	GEMMA	5.3K	Salinity stress	BLUEs, ADMIXTURE, R-programming	<i>RPLP0</i> , <i>EMB8-like</i>	Ca2, Ca4	Statistical analysis, Molecular markers, Cross-validations	(Ahmed et al., 2021)
Chickpea accessions	92	Winter	CMLM, EMMAX	16.59K	Zn and Fe concentration in seeds	GAPIT, STACKS v1.0, FASTQC v0.10.1, CGAP v1.0, SnpEff v3.1h, BiNGO plugin of Cytoscape V2.6, PAMLv4.8a, TASSEL v5.0, PowerMarker v3.51, MEGA v5.0, STRUCTURE v2.3.4, PLINK, MALDI-TOF	Ca03227, Ca03400, Ca24399, Ca22196, Ca09146, Ca03842, Ca00947, Ca12262, Ca09416, Ca27126, Ca19289, Ca06927	Chr.1, Chr.2, Chr.3, Chr.4, Chr.5, Chr.7	Association mapping, QTL Mapping, HPLC, Molecular markers, RT-PCR, Statistical analysis	(Upadhyaya et al., 2016)

will not be expressed. Additionally, if the allele expression is being influenced by environmental factors (soil, light, temperature, etc.) then the dominant trait may only emerge under certain environmental conditions. Thus, phenotype is the sum of three-dimensional (3D) spatiotemporal expression information resulting from interactions between environmental factors and genotype. However, the acquisition of phenotypic data is still a bottleneck limiting functional genomics studies (Deery et al., 2016). Traditional phenotypic approaches mostly depend on manual measurements, which are subjective, time-consuming, laborious, and hamper comprehensive phenotypic data from individuals within a large population. Additionally, errors are obvious in manual measurements, and therefore, data reliability and accuracy data cannot be guaranteed (Xiao et al., 2022). In addition to cost, manpower, and other related limitations, manual measurements can only be exploited for limited features during the critical stages of plant growth. Moreover, physical changes cannot be fully detected throughout a plant's life cycle. The aforementioned shortcomings and limitations from traditional approaches can be overcome by exploiting high throughput phenotyping (HTP). HTP is emerging as an important tool for evaluating a plant's phenotype. HTP approaches such as fluorescence imaging, hyperspectral imaging, visible light imaging, automation technology, machine vision and advanced sensors combined with advanced information technologies (ITs) and data extraction systems have enabled more accurate, rapid, and non-destructive measurements of physiological and morphological parameters. Each of the above-mentioned techniques has its advantages that allow reliability and accuracy in high throughput detection (Jiang et al., 2018; Narisetti et al., 2021; Sarkar et al., 2021).

HTP platforms integrate data acquisition equipment, a control terminal, and data analysis platforms. Firstly, in HTP, phenotypic data are collected via spectroscopy and non-invasive imaging techniques and then high-performance computational tools are adopted to rapidly analyze plant physiological state and other growth activities. In comparison to traditional phenotypic approaches, HTP offers simultaneous data acquisition of multiple traits and close observation of plant activities at different growth stages throughout the life cycle. Secondly, traditional approaches like visual scoring, are prone to subjective interpretation while trait characterization in HTP is more based on images or spectra which are more objective. Thirdly, HTP offers modeling-based non-destructive estimation of biochemical parameters, hence reducing laborious tasks and time. In the last few years, there have been major advances in HTP techniques to study different targets such as plant roots, leaves, shoots, seeds, cells, and canopy (Yang et al., 2020). For example, microscopic imaging and microcomputed tomography (m-CT) are used in the determination of tissue morphology (Zhang et al., 2021), cell growth rate (Gallegos et al., 2020), alterations in cell structure (Faulkner et al., 2017) and number of cells (Mele and Gargiulo, 2020). Moreover, visible light imaging and 3D graphics have intensively been used for characterization of seed morphological traits like germination rate (Ligterink and Hilhorst, 2017; Merieux et al., 2021), seed weight (Huang et al., 2022), growth and development (Margapuri et al., 2021), coleoptiles length (Zhang and Zhang, 2018) and seed color

(Baek et al., 2020). Other physiological, morphological, and biochemical parameters have also been intensively studied through combined GWAS and HTP using time domain pulsed nuclear magnetic resonance (NMR) (Melchinger et al., 2018), Semantic Guided Interactive Object Segmentation (SGIOS) (Yuan et al., 2022), Graphical User Interface (GUI) (Yuan et al., 2022), Near-infrared spectroscopy (Jasinski et al., 2016; Anderson et al., 2019a), Deep convolutional neural networks (DCNNs) (Jiang et al., 2021), Hyper-spectral vegetation indices (VIs) (Koh et al., 2022), unmanned aerial vehicle (UAV) (Jiang et al., 2021), computed tomography (Guo et al., 2022) and multi-spectral or hyper-spectral images (Wu et al., 2021a; Correia et al., 2022). In-depth information on phenotyping techniques can be found here (Rahaman et al., 2015). Zhang and Zhang (2018), in their review, summarized the applications of recently developed imaging HTP techniques to study the pathological, physiological, and morphological traits of plants (Zhang and Zhang, 2018). Shakoore et al. (2017), provided a detailed review of HTP techniques (especially recently developed sensors) in accelerating plant breeding and disease assessments (Shakoore et al., 2017). Recently Liu et al. (2020), thoroughly reviewed hyper-spectral imaging and three dimensional (3D) techniques applications for plant phenotyping (Liu et al., 2020). Jang et al. (2020), in their review have focused on UAV applications in plant breeding and summarized the deployed sensors that can be mounted on UAV and their characteristics in detail (Jang et al., 2020).

Phenotypic data is one of the most important factors limiting GWAS power, inaccurate and non-reliable phenotypic data results in false associations. For example, imprecise phenotypic data greatly influence the true MAF present within a population, so that the identified SNPs cannot be linked to traits that are affected by these SNPs. Phenotypic data collected manually is always prone to error. Therefore, to minimize these errors, HTP techniques are combined with GWAS. The success of this combination can be gauged by the number of studies published in the last 4 years. HTP combination with GWAS has made it possible to study those plant traits that cannot be studied through physical phenotypic parameters e.g. I-traits (traits that can only be studied efficiently through images) (Wu et al., 2021a). Furthermore, this combination also improves the crop selection process and makes selection strategies tractable for plant breeders to increase the rate of genetic gain (Crain et al., 2018). Wu et al. (2021a), combined an HTP technique called Plant array, a lysimetric-based system developed by Halperin et al. (2017), which combines several factors to measure plant water relations during plant life cycle with GWAS to study the physiological parameters for drought stress in 106 accessions of cowpea (Halperin et al., 2017; Wu et al., 2021b). They identified a total of 20 SNPs out of which 14 were significantly associated with critical soil water content (θ_{crit}) and 6 were significantly associated with the slope of transpiration rate declining (K_{Tr}). The detected SNPs were distributed on 9 different chromosomes and accounted for 8.7 to 21% of phenotypic variation, indicating both stomatal closure speed and stomatal sensitivity to soil drought were controlled by multiple genes with moderate effects. Wu et al. (2021b) established a multi-optical HTP system based on X-ray computed tomography and hyper-spectral imaging combined with GWAS to study drought

stress in 368 maize genotypes using an I-trait pipeline (Wu et al., 2021b). Their data revealed 4322 significant locus-trait associations, representing 1529 QTLs and 2318 candidate genes. They also reported two novel genes *ZmFAB1A* and *ZmCPGM2* associated with drought stress and 15 I-traits as potential markers for maize drought tolerance breeding. Crain et al. (2022) combined the unmanned aerial vehicle (UAV) HTP technique with GWAS to study the relationships between single plant and full plot yield in 340 wheat accessions using association mapping panel for full plot and single plant association mapping for single plants (Crain et al., 2022). UAV (equipped with a multi-spectral camera) was used to collect normalized difference vegetation index (NDVI) throughout seasons (2018-2019 and 2019-2020). According to their data, both single plant and full plot NDVI measurements (during the grain filling stage) were positively associated with grain yield. They identified SNPs on chromosome 7A and 2B significantly associated with spikelet and spike length, respectively, during the growing season 2018-2019 but with no associations for the same traits were identified in 2019-2020 growing season. Moreover, SNPs marker identified on chromosome 4B were significantly associated with plant height within the full plot association mapping panel in both seasons. However, no association was found for the same trait within single plant association mapping for a single plant. Furthermore, canopy reflectance spectrometry combined with GWAS in strawberries increased the selection efficiency of resistant lines against powdery mildew (Tapia et al., 2022). Aerial-based systems combined with GWAS have greatly facilitated the measurements of canopy traits such as canopy coverage and lodging to further facilitate the identification of novel QTLs associated with such traits. RGB (Red, Green, and Blue) imaging and GWAS combination have been successfully exploited in detecting the genetic architecture related to disease resistance. Silva et al. (2022) used a ground-based proximal sensing HTP platform in combination with a DJI quadcopter Matric-100 multi-spectral imaging camera to screen wheat genotypes against barley yellow dwarf disease (BYD) (Silva et al., 2022). GWAS analysis identified 16 significant SNPs marker associated with resistance to BYD distributed on chromosomes 5AS, 7AL, and 7DL. They also identified the *Bdv2* gene on chromosome 7AL as having a strong association with resistance to BYD. Xiao et al. (2022) provided a review of advances in HTP techniques and also summarized the combined applications of HTP and GWAS in different crops such as wheat, rice, barley, maize, soybean, and other species till 2020 (Xiao et al., 2022).

So far, no study has been reported on a combined analysis of HTP and GWAS in mung bean. This combination of HTP and GWAS in mung bean can be useful for studying novel traits such as i-traits associated with biotic and abiotic stresses (Guo et al., 2018). Such traits can only be efficiently measured or calculated through aerial or imaging techniques. X-ray computed tomography; multi-spectral imaging, spectroscopy, 3D structural analysis, and RGB imaging can be used in mung bean to study the physiological and biochemical activities under stressful conditions throughout the life cycle. Combining the aforementioned HTP techniques with GWAS can identify novel loci or genes associated with yield-related traits and resistance to biotic and abiotic stresses. HTP techniques can

measure phenotypic traits more rapidly and accurately and also improve selection efficiency in mung bean breeding programs.

5 Connecting GWAS with genome editing

Genome editing (GE) technologies have revolutionized the field of life science by precisely editing plant genomes. In the past few years, different GE tools such as zinc finger nucleases (ZFNs), transcriptional activator-like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeat (CRISPR) have been successfully exploited for editing complex and simple plant traits. ZFNs are targetable DNA cleavage proteins that act as restriction enzymes to cut DNA sequences. ZFNs were artificially developed by fusing binding domains of ZFNs proteins with the Fok-I endonuclease cleavage domain. Similarly, TALENs were also developed by fusing TALEs (transcription activator-like effectors) derived DNA binding domains with the Fok-I endonuclease cleavage domain (Zhang et al., 2019). TALENs are capable of inducing double-stranded breaks (DSBs) in targeted sequences, which activates DNA repair pathways, resulting in genome modifications. However, both TALENs and ZFNs have been intensively used to edit the genome of living organisms including humans and plants, but some limitations of these technologies have prevented their effective use. Therefore, scientists started looking for other effective GE technologies and discovered the CRISPR-Cas9 system in archaea and bacteria (Jinek et al., 2012) (Figure 5). In the beginning, CRISPR-Cas also had limitations just like other GE technologies, but with time, different CRISPR-Cas variants were discovered to overcome these limitations. CjCas9 is a Cas9 variant, derived from *Campylobacter jejuni*, and is more specific in cutting targeted DNA sequences than Cas9 *in vivo* and *in vitro*. CjCas9 is delivered through AAV (adeno-associated virus) in the target cell and induces targeted mutations at high frequency (Kim et al., 2017). Recently discovered Cas13 is another variant that is used to target endogenous RNAs and viral RNAs in plant cells (Wolter and Puchta, 2018). Different research groups have reported that CRISPR-Cas13 is highly efficient and has the highest RNA target specificity compared with other Cas variants (Abudayyeh et al., 2017). NGS technologies have made precise target-specific gene editing much easier. Significantly associated SNPs controlling important traits have made CRISPR-Cas base editing more efficient than whole gene insertion and deletion. Combining GWAS and CRISPR-Cas system offers three key advantages; firstly, editing of identified SNPs/genes with CRISPR can further validate whether the identified SNPs/genes are indeed associated with trait of interest or not, secondly, putative genes with unknown functions identified through GWAS can be knocked-out to identify their functions, thirdly, insertion or deletion in candidate gene (identified through GWAS) can help in improving plant traits. For example, Kariyawasam et al. (2022) identified *SnTox5* (involved in facilitating *parastagonospora nodorum* colonization in mesophyll tissue of wheat to induce program cell death) gene using GWAS and edited through CRISPR-Cas system to further validate its previously reported role in pathogenesis. They identified *Sn2000_06735* (putative candidate gene) as a homolog of *SnTox5* and to validate

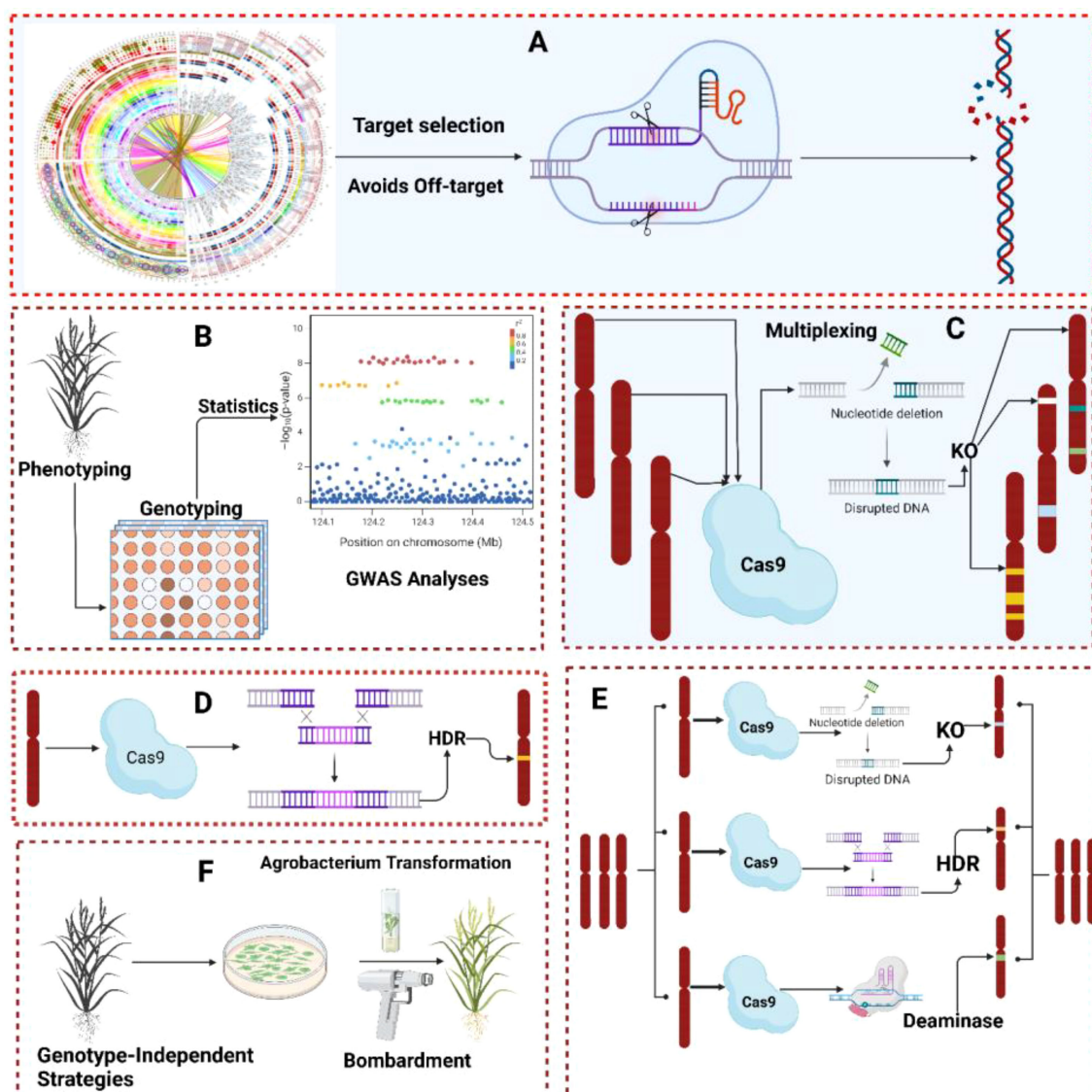


FIGURE 5

A simultaneous representation of GWAS and genome editing. (A) General overview of CRISPR-Cas from gene selection to genome editing. (B) Phenotyping, genotyping, and identification of the causal loci(s)/allele(s) associated with particulate trait. (C) Genome editing of loci/alleles identified by GWAS for further validation of results using gene knockout strategy (D) Genome editing of loci/alleles identified by GWAS for further validation of results using gene HDR and NHEJ strategy (E) Genome editing of loci/alleles identified by GWAS for further validation of results using gene KO, HDR, NHEJ and deaminase strategy (F) CRISPR-Cas most reliable delivery methods (Agrobacterium and Bombardment).

this, *Sn2000_06735* was disrupted by inserting *hyg^R* (hygromycin resistance cassette) using the CRISPR-Cas system. *Sn2000_06735* disrupted mutants failed to cause necrosis and prevented *Parastagonospora nodorum* colonization (Kariyawasam et al., 2022).

Thus, confirming *Sn2000_06735* is associated with *Parastagonospora nodorum* pathogenesis. Similarly, Liu et al. (2021) identified the *Fov7* gene (encodes for GLR proteins) through GWAS that is associated with resistance to *Fusarium oxysporum* in *Gossypium hirsutum*. CRISPR-Cas system-based knockout of *Fov7* resulted in extreme susceptibility to *Fusarium oxysporum* in all-cotton lines. Moreover, they also identified the

significant SNP in the *Fov7* gene associated with resistance to *Fusarium oxysporum* and revealed that this SNP changes an amino acid and confers resistance. Another group of researchers selected different rice cultivars using pedigree analysis to identify yield-related candidate genes through GWAS. They discovered six genes with known functions (associated with yield) and 123 loci with genes of unknown functions. From 123 loci, they randomly selected 57 genes for CRISPR-Cas-based system knock-out to identify their functions. Their results revealed that most of these genes were significantly associated with yield-related traits. For instance, *Os1g0885000*, *Os1g088600*, and *Os1g0555100* showed

fewer tillers, a reduction in plant growth, and changes in panicle structure, respectively (Huang et al., 2018). Liang et al. (2022) phenotyped 2409 accessions of soybeans to identify the candidate gene involved in controlling the number of branches per plant (Liang et al., 2022). GWAS analysis revealed *SoyZH13_18g242900* (also known as *Dt2*) as a candidate gene significantly associated with the increase in number of branches per plant and several other agronomic traits. To validate the role of *SoyZH13_18g242900*, DN50 (soybean variety with four branches) was selected and *SoyZH13_18g242900* was knocked out using the CRISPR-Cas9 system. Field experiments revealed that *Dt2* mutant lines showed an increase in the number of branches compared with wild-type DN50. Moreover, these mutant lines also increased days to flowering and maturity and enhanced the number of nodes per plant and plant height.

6 Future prospects

6.1 Opportunities, challenges, and future strategies of GWAS and PWAS

The prior knowledge of natural genetic variations present in mung bean is extending and making mung bean a model crop to study genetic variations in other crops like mash bean, faba bean, and other pulses. We have observed these advancements in recent years through a large number of genetic variability studies conducted to understand the phenomena of natural variation in mung beans. GWAS soon will be more useful/informative in mung bean using advanced sequencing technologies to unlock the hidden genetic variations and availability of the high throughput SNPs set associated with phenotype as a reference genome in genebank e.g., IPK, to study the mutations in mung bean mutant genotypes and construct some useful genetic maps such as MutMap. The output of GWAS could be executed and utilized in different aspects, for example, improving breeding programs, targeted genome editing, identification of novel genes, constructing genetic maps, high throughput phenotyping or highly accurate phenotyping by breeders can also improve GWAS power in detecting new loci and recombinations. These advances help in facilitating and improving breeding by analyzing the genomics or genetics of agronomically important plant traits. In-depth analysis in detecting causative loci via GWAS, for instance, haplotype-based analysis is a key for genomics-assisted plant breeding. In comparison to QTLs mapping, GWAS has higher resolution due to the large number of recombination's and large population comprising hundreds to thousands of genotypes used to study genetic variations in more depth and breadth. GWAS in future mung bean work must be considered as an exploratory analysis for selecting true segregating parents which can be utilized in developing populations and QTL mapping and in the future for molecular and genetic association validations. Besides, GWAS is also useful in understanding marker-based selection (individual selection for breeding programs based on their available genetic

information of specific alleles linked to QTLs) or breeding-program-based variation (the genetic variability of association panel implemented in improving crops) because the association mapping population is considered as a source of alleles that are rarely present in bi-parental mapping populations. Recently, various studies used both association mapping and QTLs mapping to isolate or identify and validate the QTLs associated with traits of interest for example, brassica (He et al., 2017), maize (Zhao et al., 2018b) and faba bean (Sallam et al., 2016). This technique utilizes both populations (biparental and mixed population) to determine whether the identified significant markers are associated with the same trait of interest in two different genetic backgrounds or not. However, no study in mung bean has been reported yet using this technique and therefore, it will be of great advantage to implement it in mung bean to genetically improve the traits of interest. Association mapping population is always rich in alleles (including land races, wild types and domestication alleles) and offers great genetic variation; therefore, it can be considered as an excellent genetic resource and enhance the chances of discovering new genes/alleles controlling complex traits such as yield, tolerance to biotic and abiotic stresses. The analyses enable predicting the function(s) of the different alleles representing genetic alterations/mutations and candidate alleles/genes which are associated or have an agronomic impact, thus could be utilized in molecular validations such as genome editing and gene expression. Collaborations with bioinformaticians and statisticians can help in establishing new efficient statistical models and databases that can be utilized during the analysis of complex traits. Integration of genetics and omics can be crucial for molecular analysis. Therefore, they should be integrated and implemented together. The expansion in natural variation analysis to molecular mechanisms will further provide insights into mechanisms involved in mung bean growth, adaptation, and development.

The advancements in genomic approaches offer opportunities to characterize genetic diversity, traits mapping, and improvements and they also offer a greater understanding of complex genomes and the development of new genome editing tools for breeding.

6.1.1 Complex polyploid genome, genetic resources, and rapid domestication of crop species

Autopolyploidy and allopolyploidy are common mechanisms of genome doubling and many plants (especially angiosperms) during evolution have undergone at least two rounds of polyploidy. This natural mechanism results in introducing more allelic diversity, improving crop adaptation to new environmental conditions and new phenotypic variations. Plant breeders have already taken several advantages of this mechanism by introducing artificial polyploids with an increase in fruit size (Wu et al., 2012), developing seedless fruits (Varoquaux et al., 2000), and increasing the grain yield (Rosyara et al., 2019). Genomic studies in polyploidy species have always been a great challenge due to several complications and reasons. Besides, the development of a

genomic library with high quality, there is another challenge due to the inclusion of different but closely related sub-genomes, differentiating homologous loci and generating non-mosaic sub-genome scaffolds. Different research groups have made efforts to reduce the genomic complexity of polyploids by sequencing closely related species (Shulaev et al., 2011) or diploid progenitors (D'hont et al., 2012) to generate initial reliable reference assemblies. Detection of SVs and SNPs in closely related species is still very challenging and difficult and most of the studies have failed in detecting these variations (Gordon et al., 2020). Besides these difficulties, genetic improvement of polyploids is subject to further complications: (1) dissecting the genetic architecture of complex traits becomes impossible when the variants are not mapped to the correct sub-genome (Ramírez-González et al., 2018) and (2) biologically, the exact prediction of phenotype based on genotype might be hampered by extensive epistatic interactions and regulatory feedback between sub-genomes in polyploids (Bird et al., 2018). However, these issues have already been addressed through advancements in sequencing and assembly algorithms. As the numbers of GWAS and Pan-genomic studies are expanding in polyploid crop species, we expect that the degree of SNPs, *k-mers*, and SVs will be greater compared with diploid species.

Breeding efforts using pan-genomic studies are limited because only a few research groups are using this technique and therefore, the genomic resources remain low. For example, *Silphium integrifolium* (an oil crop species with large genome size) genome was studied using transcriptome assemblies to identify loci associated with adaptation in different climatic conditions due to the non-availability of whole genome reference genomic assembly (Raduski et al., 2021). SVs remained uncharacterized in this study due to limited genomic resources and SNPs helped in identifying the loci by re-sequencing. Forage crops and turfgrass are other examples of crops with limited genomic resources. GWAS and PWAS have unlocked the challenges associated with crop domestication, especially with the reduction in the time frame generally required for developing a single variety. Plant breeders can use genomic information resulting from GWAS/PWAS to genetically improve crops efficiently by genome editing techniques or identifying markers or variants (PAV, CNV, and SNPs) associated with particular traits in wild plants. For instance, pan-genomic in tomatoes revealed that variations in fruit size/weight are controlled by the duplication of the *SKILUH* (cytochrome P450) gene (Alonge et al., 2020), rather than an SNP as reported earlier (Chakrabarti et al., 2013). Later, this was confirmed by using CRISPR-Cas9 to reduce the *SKILUH* copy number, and resulted in alterations in fruit weight (Alonge et al., 2020). Domestication of crops has significantly reduced the genetic diversity compared with wild relatives. Identification and utilization of the genetic diversity from wild relatives is a major focus of a plant breeder in improving crops. Combined applications of GWAS and genome editing technologies will allow *de-novo* domestication of wild plants and take advantage of available genetic diversity from secondary and tertiary gene pools (wild plants). Wild relatives of mung bean are known to possess high genetic diversity. Therefore, domestication with wild relatives is easy as till now no study has reported the

combining ability barriers between domesticated and wild parents. For instance, the mini core of mung bean from world vegetable gene bank Taiwan, studied GWAS in a large mung bean population (containing all the domesticated and wild relatives) to identify the SNPs associated with the trait of ineptest. They identified several SNPs associated with the trait of interest are in wild relatives rather than in the domesticated plants. The wild relatives had more SNPs and had more strong association with phenotypic traits (Sokolkova et al., 2020). This study is the proof that the domestication has reduced the genetic diversity in the mung bean to significant level. However, this can be restored by crossing the domesticated mung bean plants back to wild relatives.

6.2 Challenges, future applications, and role of high throughput phenotyping in GWAS

Various studies have demonstrated the potential applications and role of HTP in plant research but few studies have integrated GWAS and HTP. The key factors that limit GWAS and HTP integration are challenges in the accession of genomic data, accuracy in characterizing phenotypic traits, and shortage of skilled persons. Genomic data can be obtained from re-sequencing or the gene banks. Reduction in the cost of whole genome re-sequencing has made genomic studies easier but it is still time-consuming and highly laborious. On the other hand, the data available in the genebanks at the movement may not match the actual samples due shortage of genebanks. Highly accurate phenotypic data of any trait is necessary for GWAS but currently available HTP techniques applied in GWAS are still generally flawed. Many HTP techniques such as X-ray CT, hyper-spectral imaging, and visible light/RGB imaging, strongly rely on data/image processing algorithms. Recently, signal-based algorithms have been associated with several deficiencies like inaccurate feature extraction, imperfection, and low efficiency; therefore, they need to be subjected to required improvements. Highly sensitive and high-resolution equipment utilized for fluorescence imaging, X-ray CT, and hyper-spectral imaging are very expensive and therefore cannot be implemented extensively. UAVs for near-surface HTP are appropriate for collecting phenotypic canopy data in the field due to wide spatial coverage and flexibility. However, complex approaches, huge prices, and insufficient payload acquired for processing enormous remote sensing data may limit their adaptation. Besides these, there is a need to introduce more promising HTP techniques like optical coherence tomography and infrared-thermal imaging in GWAS. Currently, some efforts have already been taken to acquisition of highly accurate phenotypic data using currently available HTP techniques (Mochida et al., 2019; Liu et al., 2020; Yang et al., 2020). Changes in phenotypic data are due to alterations in genetic composition and environmental factors. Environmental changes are directly associated with changes in the phenotypic traits of plants which are difficult to control. Indeed, we can develop phenotypic databases through HTP techniques but only if we consider a wide range of environmental situations which is challenging.

To enhance the implementation of HTP in GWAS to explore the underlying complex genetic architecture of phenotypic traits in mung bean and other plant species, the following aspects must be considered,

1. Enhance the amount of investment in developing more efficient and highly accurate population genotypic data approaches.
2. Sufficient genotypic data of various crops including mung bean must be present in online databases (<https://bigd.big.ac.cn/gvm/home>) and must be accessed by all the researchers. Collection of plant material of known genotypes for HTP is one of the potential strategies to reduce the associated cost.
3. The development of efficient HTP techniques with low cost is another strategy to encourage the wider application and adaptation of this technique in GWAS.
4. The development of new and improvement of existing public phenotypic databases are of great interest to efficiently resolve resource issues in data provision, heterogeneous data formats, and insufficient meta-data. We strongly recommend and urge the publication of meta-data, that need to be structured according to the principles of FAIR (Wilkinson et al., 2016) and state the detailed information like environmental conditions and data formats. Ćwiek-Kupczyńska et al. (2016) have already proposed the guidelines for governing the description of phenotypic data, which provided a document of Minimum Information About a Plant Phenotyping Experiment (MIAPEE) and encouraged to implementation of ISA-Tab format for meta-data set organization (Ćwiek-Kupczyńska et al., 2016). Furthermore, we need to develop universal unified standard formats for phenotypic data recorded using different approaches. Some efforts are under the mission of generating efficient phenotypic databases like PHENOPSIS DB for *Arabidopsis thaliana* (<http://bioweb.supagro.inra.fr/phenopsis/>). This database can be used as a template to develop more phenotypic databases for other crops like mung bean including other pulses and cereals.
5. Incessant developments of imaging algorithms or multivariate data are essential. For example, image processing and voluminous data in-depth processing have shown excellent impact in understanding the data, owing to their unique strength in the form of self-learning ability and efficiency in large data analysis. We do not doubt that in the future the applications of in-depth learning considering plant traits data extraction will be a hot research topic.
6. A combination of all existing HTP techniques might greatly facilitate the evaluation of plant traits in different aspects.

We are urgently in need a large number of studies implementing GWAS and HTP techniques together to study diverse plant populations and traits to understand functional genetics/genomes in greater depth.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary Material](#). Further inquiries can be directed to the corresponding author.

Author contributions

SA: Conceptualization, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. MA: Formal analysis, Supervision, Validation, Writing – review & editing. AH: Conceptualization, Supervision, Validation, Writing – review & editing. MG: Data curation, Formal analysis, Software, Writing – original draft, Writing – review & editing. MS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

We are indebted to give appreciation to the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, and the Department of Biological Sciences, Nuclear Institute for Agriculture and Biology (NIAB), Faisalabad, for providing an umbrella for the research program.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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.2024.1436532/full#supplementary-material>

References

- Abou-Khater, L., Maalouf, F., Jighly, A., Alsamman, A. M., Rubiales, D., Risipail, N., et al. (2022). Genomic regions associated with herbicide tolerance in a worldwide faba bean (*Vicia faba* L.) collection. *Sci. Rep.* 12, 158. doi: 10.1038/s41598-021-03861-0
- Abraham, G., Qiu, Y., and Inouye, M. (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* 33, 2776–2778. doi: 10.1093/bioinformatics/btx299
- Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., et al. (2017). RNA targeting with CRISPR-cas13. *Nature* 550, 280–284. doi: 10.1038/nature24049
- Ahmed, S. M., Alsamman, A. M., Jighly, A., Mubarak, M. H., Al-Shamaa, K., Istanbuli, T., et al. (2021). Genome-wide association analysis of chickpea germplasms differing for salinity tolerance based on DArTseq markers. *PLoS One* 16, e0260709. doi: 10.1371/journal.pone.0260709
- Ahmed, S. R., Anwar, Z., Shahbaz, U., Skalicky, M., Ijaz, A., Tariq, M. S., et al. (2023). Potential role of silicon in plants against biotic and abiotic stresses. *Silicon* 7, 3283–3303. doi: 10.1007/s12633-022-02254-w
- Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958
- Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145–161.e123. doi: 10.1016/j.cell.2020.05.021
- Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. doi: 10.1016/j.cell.2016.05.063
- Anderson, J. V., Wittenberg, A., Li, H., and Berti, M. T. (2019a). High throughput phenotyping of *Camelina sativa* seeds for crude protein, total oil, and fatty acids profile by near infrared spectroscopy. *Ind. Crops Products* 137, 501–507. doi: 10.1016/j.indcrop.2019.04.075
- Anderson, R., Edwards, D., Batley, J., and Bayer, P. E. (2019b). Genome-wide association studies in plants. *eLS* 1–7. doi: 10.1002/9780470015902.a0027950
- Azmah, U. N., Makeri, M. U., Bagirei, S. Y., and Shehu, A. B. (2023). Compositional characterization of starch, proteins and lipids of long bean, dwarf long bean, mung bean and French bean seed flours. *Measurement: Food* 12, 100111. doi: 10.1016/j.meafoo.2023.100111
- Baek, J., Lee, E., Kim, N., Kim, S. L., Choi, I., Ji, H., et al. (2020). High throughput phenotyping for various traits on soybean seeds using image analysis. *Sensors* 20, 248. doi: 10.3390/s20010248
- Band, G., and Marchini, J. (2018). BGEN: a binary file format for imputed genotype and haplotype data. *BioRxiv* 308296. doi: 10.1101/308296
- Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., and Edwards, D. (2020). Plant pan-genomes are the new reference. *Nat. Plants* 6, 914–920. doi: 10.1038/s41477-020-0733-0
- Bayer, P. E., Golicz, A. A., Tirnaz, S., Chan, C. K. K., Edwards, D., and Batley, J. (2019). Variation in abundance of predicted resistance genes in the *Brassica oleracea* pan-genome. *Plant Biotechnol. J.* 17, 789–800. doi: 10.1111/pbi.2019.17.issue-4
- Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S., and Pirinen, M. (2016). FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501. doi: 10.1093/bioinformatics/btw018
- Bird, K. A., Vanburen, R., Puzey, J. R., and Edger, P. P. (2018). The causes and consequences of subgenome dominance in hybrids and recent polyploids. *New Phytol.* 220, 87–93. doi: 10.1111/nph.2018.220.issue-1
- Breria, C. M., Hsieh, C. H., Yen, J.-Y., Nair, R., Lin, C.-Y., Huang, S.-M., et al. (2020b). Population structure of the world vegetable center mungbean mini core collection and genome-wide association mapping of loci associated with variation of seed coat luster. *Trop. Plant Biol.* 13, 1–12. doi: 10.1007/s12042-019-09236-0
- Breria, C. M., Hsieh, C.-H., Yen, T.-B., Yen, J.-Y., Noble, T. J., and Schaffeltnier, R. (2020a). A SNP-based genome-wide association study to mine genetic loci associated to salinity tolerance in mungbean (*Vigna radiata* L.). *Genes* 11, 759. doi: 10.3390/genes11070759
- Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015
- Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Consortium, S.W.G.O.T.P.G., et al. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* 47, 291–295. doi: 10.1038/ng.3211
- Burgess, S., Scott, R. A., Timpson, N. J., Davey Smith, G., Thompson, S. G., and Consortium, E.-I. (2015). Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* 30, 543–552. doi: 10.1007/s10654-015-0011-z
- Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z
- Chakrabarti, M., Zhang, N., Sauvage, C., Muñoz, S., Blanca, J., Cañizares, J., et al. (2013). A cytochrome P450 regulates a domestication trait in cultivated tomato. *Proc. Natl. Acad. Sci.* 110, 17125–17130. doi: 10.1073/pnas.1307313110
- Chiteri, K. O., Jubery, T. Z., Dutta, S., Ganapathysubramanian, B., Cannon, S., and Singh, A. (2022). Dissecting the root phenotypic and genotypic variability of the iowa mung bean diversity panel. *Front. Plant Sci.* 12, 808001. doi: 10.3389/fpls.2021.808001
- Correia, P. M., Cairo Westergaard, J., Bernardes Da Silva, A., Roitsch, T., Carmo-Silva, E., and Marques Da Silva, J. (2022). High-throughput phenotyping of physiological traits for wheat resilience to high temperature and drought stress. *J. Exp. Bot.* 73, 5235–5251. doi: 10.1093/jxb/erac160
- Crain, J., Mondal, S., Rutkoski, J., Singh, R. P., and Poland, J. (2018). Combining high-throughput phenotyping and genomic information to increase prediction and selection accuracy in wheat breeding. *Plant Genome* 11, 170043. doi: 10.3835/plantgenome2017.05.0043
- Crain, J., Wang, X., Evers, B., and Poland, J. (2022). Evaluation of field-based single plant phenotyping for wheat breeding. *Plant Phenome J.* 5, e20045. doi: 10.1002/ppj2.20045
- Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., et al. (2016). Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods* 12, 1–18. doi: 10.1186/s13007-016-0144-4
- D'hont, A., Denoeud, F., Aury, J.-M., Baurès, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature* 488, 213–217. doi: 10.1038/nature11241
- Daovongdeuan, S., Sanyawon, D., Kim, M. Y., Jeong, H., and Lee, S. H. (2017). GWAS for seed color and size in mungbean (*Vigna radiata* (L.) Wilczek), in *Proceedings of the Korean society of crop science conference: the Korean society of crop science* (Seoul, Republic of Korea: The Korean Society of Crop Sci), 274–274.
- Das, S., Porter, L. D., Ma, Y., Coyne, C. J., Chaves-Cordoba, B., and Naidu, R. A. (2022). Resistance in lentil (*Lens culinaris*) genetic resources to the pea aphid (*Acyrtosiphon pisum*). *Entomologia Experimentalis Applicata* 170, 755–769. doi: 10.1111/eea.v170.8
- Davis, B. D. (1949). The isolation of biochemically deficient mutants of bacteria by means of penicillin. *Proc. Natl. Acad. Sci.* 35, 1–10. doi: 10.1073/pnas.35.1.1
- Deery, D. M., Rebetzke, G. J., Jimenez-Berni, J. A., James, R. A., Condon, A. G., Bovill, W. D., et al. (2016). Methodology for high-throughput field phenotyping of canopy temperature using airborne thermography. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01808
- Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., and Dermizakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* 8, 15452. doi: 10.1038/ncomms15452
- De Leeuw, C. A., Mooij, J. M., Heskes, T., and Posthuma, D. (2015). MAGMA. *PLoS Computational Biology* 11. doi: 10.1371/journal.pcbi.1004219
- Duncan, L. E., Ostacher, M., and Ballon, J. (2019). How genome-wide association studies (GWAS) made traditional candidate gene studies obsolete. *Neuropsychopharmacology* 44, 1518–1523. doi: 10.1038/s41386-019-0389-5
- Faulkner, C., Zhou, J., Evrard, A., Bourdais, G., Maclean, D., Häweker, H., et al. (2017). An automated quantitative image analysis tool for the identification of microtubule patterns in plants. *Traffic* 18, 683–693. doi: 10.1111/tra.2017.18.issue-10
- Fedoruk, M. (2013). *Linkage and association mapping of seed size and shape in lentil* (Saskatoon, Canada: University of Saskatchewan).
- Flint-García, S. A., Thornsberry, J. M., and Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* 54, 357–374. doi: 10.1146/annurev.arplant.54.031902.134907
- Gallegos, J. E., Adames, N. R., Rogers, M. F., Kraikivski, P., Ibele, A., Nurzynski-Loth, K., et al. (2020). Genetic interactions derived from high-throughput phenotyping of 6589 yeast cell cycle mutants. *NPJ Syst. Biol. Appl.* 6, 11. doi: 10.1038/s41540-020-0134-z
- Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091–1098. doi: 10.1038/ng.3367
- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., et al. (2019). The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* 51, 1044–1051. doi: 10.1038/s41588-019-0410-2
- García-Fernández, C., Campa, A., Garzón, A. S., Miklas, P., and Ferreira, J. J. (2021). GWAS of pod morphology and color characters in common bean. *BMC Plant Biol.* 21, 1–13. doi: 10.1186/s12870-021-02967-x
- Gayacharan, A., Gupta, K., Gupta, V., Tyagi, V., and Singh, K. (2020). Mungbean genetic resources and utilization. *mungbean Genome*, 9–25. doi: 10.1007/978-3-030-20008-4
- Gela, T., Ramsay, L., Haile, T. A., Vandenberg, A., and Bett, K. (2021). Identification of anthracnose race 1 resistance loci in lentil by integrating linkage mapping and genome-wide association study. *Plant Genome* 14, e20131. doi: 10.1002/tpg2.20131
- Golicz, A. A., Bayer, P. E., Barker, G. C., Edger, P. P., Kim, H., Martinez, P. A., et al. (2016). The pan-genome of an agronomically important crop plant *Brassica oleracea*. *Nat. Commun.* 7, 13390. doi: 10.1038/ncomms13390
- Gonzalez, A., Zhao, M., Leavitt, J. M., and Lloyd, A. M. (2008). Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. *Plant J.* 53, 814–827. doi: 10.1111/j.1365-313X.2007.03373.x
- Gordon, S. P., Contreras-Moreira, B., Levy, J. J., Djamei, A., Cziedik-Eysenberg, A., Tartaglio, V. S., et al. (2020). Gradual polyploid genome evolution revealed by pan-

genomic analysis of *Brachypodium hybridum* and its diploid progenitors. *Nat. Commun.* 11. doi: 10.1038/s41467-020-17302-5

Gordon, S. P., Contreras-Moreira, B., Woods, D. P., Des Marais, D. L., Burgess, D., Shu, S., et al. (2017). Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nat. Commun.* 8. doi: 10.1038/s41467-017-02292-8

Grotzinger, A. D., Rhemtulla, M., De Vlaming, R., Ritchie, S. J., Mallard, T. T., Hill, W. D., et al. (2019). Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* 3, 513–525. doi: 10.1038/s41562-019-0566-x

Guo, Z., Yang, W., Chang, Y., Ma, X., Tu, H., Xiong, F., et al. (2018). Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice. *Mol. Plant* 11, 789–805. doi: 10.1016/j.molp.2018.03.018

Guo, S., Zhou, G., Wang, J., Lu, X., Zhao, H., Zhang, M., et al. (2022). High-throughput phenotyping accelerates the dissection of the phenotypic variation and genetic architecture of shank vascular bundles in maize (*Zea mays* L.). *Plants* 11. doi: 10.3390/plants11101339

Gupta, P. K. (2021a). GWAS for genetics of complex quantitative traits: Genome to pangenome and SNPs to SVs and k-mers. *BioEssays* 43, 2100109. doi: 10.1002/bies.202100109

Gupta, P. K. (2021b). Quantitative genetics: pan-genomes, SVs, and k-mers for GWAS. *Trends Genet.* 37, 868–871. doi: 10.1016/j.tig.2021.05.006

Halperin, O., Gebremedhin, A., Wallach, R., and Moshelion, M. (2017). High-throughput physiological phenotyping and screening system for the characterization of plant–environment interactions. *Plant J.* 89, 839–850. doi: 10.1111/tpj.2017.89.issue-4

Han, X., Li, L., Chen, H., Liu, L., Sun, L., Wang, X., et al. (2022). Resequencing of 558 Chinese mungbean landraces identifies genetic loci associated with key agronomic traits. *Frontiers in Plant Science* 13, 1043784. doi: 10.21203/rs.3.rs-1729302/v1

He, Y., Wu, D., Wei, D., Fu, Y., Cui, Y., Dong, H., et al. (2017). GWAS, QTL mapping and gene expression analyses in *Brassica napus* reveal genetic control of branching morphogenesis. *Sci. Rep.* 7, 15971. doi: 10.1038/s41598-017-15976-4

Hou, D., Yousaf, L., Xue, Y., Hu, J., Wu, J., Hu, X., et al. (2019). Mung bean (*Vigna radiata* L.): Bioactive polyphenols, polysaccharides, peptides, and health benefits. *Nutrients* 11. doi: 10.3390/nut11061238

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3: Genes Genomes Genet.* 1, 457–470. doi: 10.1534/g3.111.001198

Huang, X., and Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551. doi: 10.1146/annurev-arplant-050213-035715

Huang, J., Li, J., Zhou, J., Wang, L., Yang, S., Hurst, L. D., et al. (2018). Identifying a large number of high-yield genes in rice by pedigree analysis, whole-genome sequencing, and CRISPR-Cas9 gene knockout. *Proc. Natl. Acad. Sci.* 115, E7559–E7567. doi: 10.1073/pnas.1806110115

Huang, X., Zheng, S., and Zhu, N. (2022). High-throughput legume seed phenotyping using a handheld 3D laser scanner. *Remote Sens.* 14, 431. doi: 10.3390/rs14020431

Jaiswal, V., Gupta, S., Gahlaut, V., Muthamilarasan, M., Bandyopadhyay, T., Ramchary, N., et al. (2019). Genome-wide association study of major agronomic traits in foxtail millet (*Setaria italica* L.) using ddRAD sequencing. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-41602-6

Jang, G., Kim, J., Yu, J.-K., Kim, H.-J., Kim, Y., Kim, D.-W., et al. (2020). Cost-effective unmanned aerial vehicle (UAV) platform for field plant breeding application. *Remote Sens.* 12, 998. doi: 10.3390/rs12060998

Jasinski, S., Lécureuil, A., Durand, M., Bernard-Moulin, P., and Guerche, P. (2016). Arabidopsis seed content QTL mapping using high-throughput phenotyping: the assets of near infrared spectroscopy. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01682

Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V. S., Gundlach, H., Monat, C., et al. (2020). The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature* 588, 284–289. doi: 10.1038/s41586-020-2947-8

Jiang, Y., Li, C., Robertson, J. S., Sun, S., Xu, R., and Paterson, A. H. (2018). GPhenoVision: A ground mobile system with multi-modal imaging for field-based high throughput phenotyping of cotton. *Sci. Rep.* 8, 1–15. doi: 10.1038/s41598-018-19142-2

Jiang, Z., Tu, H., Bai, B., Yang, C., Zhao, B., Guo, Z., et al. (2021). Combining UAV-RGB high-throughput field phenotyping and genome-wide association study to reveal genetic variation of rice germplasm in dynamic response to drought stress. *New Phytol.* 232, 440–455. doi: 10.1111/nph.v232.1

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., et al. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. *Nat. Genet.* 51, 1749–1755. doi: 10.1038/s41588-019-0530-8

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *science* 337, 816–821. doi: 10.1126/science.1225829

Johannsen, W. (1911). The genotype conception of heredity. *Am. Nat.* 45, 129–159. doi: 10.1086/279202

Johnson, N., Boatwright, J. L., Bridges, W., Thavarajah, P., Kumar, S., Shippe, E., et al. (2021). Genome-wide association mapping of lentil (*Lens culinaris* Medikus) prebiotic carbohydrates toward improved human health and crop stress tolerance. *Sci. Rep.* 11, 13926. doi: 10.1038/s41598-021-93475-3

Kang, Y. J., Kim, S. K., Kim, M. Y., Lestari, P., Kim, K. H., Ha, B.-K., et al. (2014). Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* 5. doi: 10.1038/ncomms6443

Kariyawasam, G. K., Richards, J. K., Wyatt, N. A., Running, K. L., Xu, S. S., Liu, Z., et al. (2022). The *Parastagonospora nodorum* necrotrophic effector SnTox5 targets the wheat gene Snn5 and facilitates entry into the leaf mesophyll. *New Phytol.* 233, 409–426. doi: 10.1111/nph.v233.1

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., et al. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* 10, e1004722. doi: 10.1371/journal.pgen.1004722

Kim, E., Koo, T., Park, S. W., Kim, D., Kim, K., Cho, H.-Y., et al. (2017). *In vivo* genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat. Commun.* 8, 14500. doi: 10.1038/ncomms14500

Kinnersley, B., Labussiere, M., Holroyd, A., Di Stefano, A.-L., Broderick, P., Vijayakrishnan, J., et al. (2015). Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat. Commun.* 6. doi: 10.1038/ncomms9559

Koh, J. C., Banerjee, B. P., Spangenberg, G., and Kant, S. (2022). Automated hyperspectral vegetation index derivation using a hyperparameter optimisation framework for high-throughput plant phenotyping. *New Phytol.* 233, 2659–2670. doi: 10.1111/nph.v233.6

Koo, Y., and Poethig, R. S. (2021). Expression pattern analysis of three R2R3-MYB transcription factors for the production of anthocyanin in different vegetative stages of *Arabidopsis* leaves. *Appl. Biol. Chem.* 64, 1–7. doi: 10.1186/s13765-020-00584-0

Kurt, F., and Filiz, E. (2020). Subcellular iron transport genes in *Arabidopsis thaliana*: insights into iron homeostasis. *J. BioSci. Biotechnol.* 9 (1), 1–10.

Lam, M., Awasthi, S., Watson, H. J., Goldstein, J., Panagiotaropoulou, G., Trubetskoy, V., et al. (2020). RICOPIII: rapid imputation for CONsortia Pipeline. *Bioinformatics* 36, 930–933. doi: 10.1093/bioinformatics/btz633

Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* 11, 241–247. doi: 10.1038/ng1195-241

Levy-Sakin, M., Pastor, S., Mostovoy, Y., Li, L., Leung, A. K., McCaffrey, J., et al. (2019). Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* 10. doi: 10.1038/s41467-019-08992-7

Li, Y.-H., Zhou, G., Ma, J., Jiang, W., Jin, L.-G., Zhang, Z., et al. (2014). *De novo* assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* 32, 1045–1052. doi: 10.1038/nbt.2979

Liang, Q., Chen, L., Yang, X., Yang, H., Liu, S., Kou, K., et al. (2022). Natural variation of Dt2 determines branching in soybean. *Nat. Commun.* 13. doi: 10.1038/s41467-022-34153-4

Ligterink, W., and Hilhorst, H. W. (2017). High-throughput scoring of seed germination. *Plant Hormones: Methods Protoc.*, 57–72.

Lin, Y. P., Chen, H. W., Yeh, P. M., Anand, S. S., Lin, J., Li, J., et al. (2023). Demographic history and distinct selection signatures of two domestication genes in mungbean. *Plant Physiol.* 2, 1197–1212. doi: 10.1093/plphys/kiad356

Liu, H., Bruning, B., Garnett, T., and Berger, B. (2020). Hyperspectral imaging and 3D technologies for plant phenotyping: From satellite to close-range sensing. *Comput. Electron. Agric.* 175, 105621. doi: 10.1016/j.compag.2020.105621

Liu, C., Wang, Y., Peng, J., Fan, B., Xu, D., Wu, J., et al. (2022a). High-quality genome assembly and pan-genome studies facilitate genetic discovery in mung bean and its improvement. *Plant Commun.* 3, 100352. doi: 10.1016/j.xplc.2022.100352

Liu, J., Xue, C., Lin, Y., Yan, Q., Chen, J., Wu, R., et al. (2022b). Genetic analysis and identification of VtFR08, a salt tolerance-related gene in mungbean. *Gene* 836, 146658. doi: 10.1016/j.gene.2022.146658

Liu, S., Zhang, X., Xiao, S., Ma, J., Shi, W., Qin, T., et al. (2021). A single-nucleotide mutation in a GLUTAMATE RECEPTOR-LIKE gene confers resistance to Fusarium Wilt in *Gossypium hirsutum*. *Advanced Sci.* 8, 2002723. doi: 10.1002/adv.20202723

Ma, Y., Marzougui, A., Coyne, C. J., Sankaran, S., Main, D., Porter, L. D., et al. (2020). Dissecting the genetic architecture of *Aphanomyces* root rot resistance in lentil by QTL mapping and genome-wide association study. *Int. J. Mol. Sci.* 21. doi: 10.3390/ijms21062129

Maalouf, F., Abou-Khater, L., Babiker, Z., Jighly, A., Alsamman, A. M., Hu, J., et al. (2022). Genetic dissection of heat stress tolerance in faba bean (*Vicia faba* L.) using GWAS. *Plants* 11. doi: 10.3390/plants11091108

Mägi, R., and Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinf.* 11, 1–6. doi: 10.1186/1471-2105-11-288

Manjunatha, P. B., Harisha, R., Kohli, M., Naik, P. K., Sagar, S. P., Shashidhar, B. R., et al. (2023). Exploring the world of mungbean: uncovering its origins, taxonomy, genetic resources and research approaches. *Int. J. Plant Soil Sci.* 20, 614–635. doi: 10.9734/ijpss/2023/v35i203846

Manuweera, B., Mudge, J., Kahanda, L., Mumey, B., Ramaraj, T., and Cleary, A. 2019 Pangenome-wide association studies with frequented regions, in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics* (New York, NY, USA: Association for Computing Machinery (ACM)), 627–632.

Mao, T., Li, J., Wen, Z., Wu, T., Wu, C., Sun, S., et al. (2017). Association mapping of loci controlling genetic and environmental interaction of soybean flowering time under

various photo-thermal conditions. *BMC Genomics* 18, 1–17. doi: 10.1186/s12864-017-3778-3

Mardis, E. R. (2017). DNA sequencing technologies: 2006–2016. *Nat. Protoc.* 12, 213–218. doi: 10.1038/nprot.2016.182

Margapuri, V., Courtney, C., and Neilsen, M. (2021). Image processing for high-throughput phenotyping of seeds. *EPiC Ser. Computing* 75, 69–79.

Mathew, B., Léon, J., and Sillanpää, M. J. (2018). A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. *Heredity* 120, 356–368. doi: 10.1038/s41437-017-0023-4

Mathivathana, M. K., Murukarthick, J., Karthikeyan, A., Jang, W., Dhasarathan, M., Jagadeeshselvam, N., et al. (2019). Detection of QTLs associated with mungbean yellow mosaic virus (MYMV) resistance using the interspecific cross of *Vigna radiata* × *Vigna umbellata*. *J. Appl. Genet.* 60, 255–268. doi: 10.1007/s13353-019-00506-x

Mbathchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* 53, 1097–1103. doi: 10.1038/s41588-021-00870-7

Mclaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17, 1–14. doi: 10.1186/s13059-016-0974-4

Melchinger, A., Böhm, J., Utz, H., Müller, J., Munder, S., and Mauch, F. (2018). High-throughput precision phenotyping of the oil content of single seeds of various oilseed crops. *Crop Sci.* 58, 670–678. doi: 10.2135/cropsci2017.07.0429

Mele, G., and Gargiulo, L. (2020). Automatic cell identification and counting of leaf epidermis for plant phenotyping. *MethodsX* 7, 100860. doi: 10.1016/j.mex.2020.100860

Merieux, N., Cordier, P., Wagner, M.-H., Ducournau, S., Aligon, S., Job, D., et al. (2021). ScreenSeed as a novel high throughput seed germination phenotyping method. *Sci. Rep.* 11.

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585, 79–84. doi: 10.1038/s41586-020-2547-7

Mishra, G. P., Dikshit, H. K., Tripathi, K., Aski, M. S., Pratap, A., Dasgupta, U., et al. (2022). “Mungbean breeding,” in *Fundamentals of field crop breeding*. (Springer Nature Singapore, Singapore), 1097–1149.

Mochida, K., Koda, S., Inoue, K., Hirayama, T., Tanaka, S., Nishii, R., et al. (2019). Computer vision-based phenotyping for improvement of plant productivity: a machine learning perspective. *GigaScience* 8, gyl153. doi: 10.1093/gigascience/gyl153

Nadeem, S., Riaz Ahmed, S., Luqman, T., Tan, D. K., Maryum, Z., Akhtar, K. P., et al. (2024). A comprehensive review on *Gossypium hirsutum* resistance against cotton leaf curl virus. *Front. Genet.* 15, 1306469. doi: 10.3389/fgenet.2024.1306469

Narisetti, N., Henke, M., Seiler, C., Junker, A., Ostermann, J., Altmann, T., et al. (2021). Fully-automated root image analysis (faRIA). *Sci. Rep.* 11, 1–15. doi: 10.1038/s41598-021-95480-y

Noble, T. J., Tao, Y., Mace, E. S., Williams, B., Jordan, D. R., Douglas, C. A., et al. (2018). Characterization of linkage disequilibrium and population structure in a mungbean diversity panel. *Front. Plant Sci.* 8, doi: 10.3389/fpls.2017.02102

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* 6, doi: 10.1038/ncomms6890

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Privé, F., Arbel, J., and Vilhjalmsón, B. J. (2020). LDpred2: better, faster, stronger. *Bioinformatics* 36, 5424–5431. doi: 10.1093/bioinformatics/btaa1029

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

Puspitasari, W., Allemann, B., Angra, D., Appleyard, H., Ecke, W., Möllers, C., et al. (2022). NIRS for vicine and convicine content of faba bean seed allowed GWAS to prepare for marker-assisted adjustment of seed quality of German winter faba beans. *J. Cultivated Plants* 74.

Raduski, A. R., Herman, A., Pogoda, C., Dorn, K. M., Van Tassel, D. L., Kane, N., et al. (2021). Patterns of genetic variation in a prairie wildflower, *Silphium integrifolium*, suggest a non-prairie origin and locally adaptive variation. *Am. J. Bot.* 108, 145–158. doi: 10.1002/ajb2.v108.1

Rafalski, J. A. (2010). Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13, 174–180. doi: 10.1016/j.pbi.2009.12.004

Rahaman, M. M., Chen, D., Gillani, Z., Klukas, C., and Chen, M. (2015). Advanced phenotyping and phenotype data analysis for the study of plant growth and development. *Front. Plant Sci.* 6, 619. doi: 10.3389/fpls.2015.00619

Rajendran, K., Coyne, C. J., Zheng, P., Saha, G., Main, D., Amin, N., et al. (2021). Genetic diversity and GWAS of agronomic traits using an ICARDA lentil (*Lens culinaris* Medik.) Reference Plus collection. *Plant Genet. Resour.* 19, 279–288. doi: 10.1017/S147926212100006X

Ramírez-González, R., Borrill, P., Lang, D., Harrington, S., Brinton, J., Venturini, L., et al. (2018). The transcriptional landscape of polyploid wheat. *Science* 361, eaar6089. doi: 10.1126/science.aar6089

Reddy, V. R. P., Das, S., Dikshit, H. K., Mishra, G. P., Aski, M. S., Singh, A., et al. (2021). Genetic dissection of phosphorous uptake and utilization efficiency traits using GWAS in mungbean. *Agronomy* 11. doi: 10.3390/agronomy11071401

Rosyara, U., Kishii, M., Payne, T., Sansaloni, C. P., Singh, R. P., Braun, H.-J., et al. (2019). Genetic contribution of synthetic hexaploid wheat to CIMMYT's spring bread wheat breeding germplasm. *Sci. Rep.* 9, 1–11. doi: 10.1038/s41598-019-47936-5

Ruperao, P., Thirunavukkarasu, N., Gandham, P., Selvanayagam, S., Govindaraj, M., Nebie, B., et al. (2021). Sorghum pan-genome explores the functional utility for genomic-assisted breeding to accelerate the genetic gain. *Front. Plant Sci.* 963. doi: 10.3389/fpls.2021.666342

Sallam, A., Arbaoui, M., El-Esawi, M., Abshire, N., and Martsch, R. (2016). Identification and verification of QTL associated with frost tolerance using linkage mapping and GWAS in winter faba bean. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.01098

Sandhu, K., and Singh, A. (2021). Strategies for the utilization of the USDA mung bean germplasm collection for breeding outcomes. *Crop Sci.* 61, 422–442. doi: 10.1002/csc2.20322

Sarkar, S., Cazenave, A.-B., Oakes, J., McCall, D., Thomason, W., Abbott, L., et al. (2021). Aerial high-throughput phenotyping of peanut leaf area index and lateral growth. *Sci. Rep.* 11, 21661. doi: 10.1038/s41598-021-00936-w

Schreinemachers, P., Ebert, A. W., and Wu, M.-H. (2014). Costing the *ex situ* conservation of plant genetic resources at AVRDC–The World Vegetable Center. *Genet. Resour. Crop Evol.* 61, 757–773. doi: 10.1007/s10722-013-0070-5

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *science* 316, 1341–1345. doi: 10.1126/science.1142382

Sehrawat, N., Yadav, M., Sharma, A. K., Sharma, V., Chandran, D., Chakraborty, S., et al. (2024). Dietary mung bean as promising food for human health: gut microbiota modulation and insight into factors, regulation, mechanisms and therapeutics—an update. *Food Sci. Biotechnol.* 1–11. doi: 10.1007/s10068-023-01495-8

Semagn, K., Björnstad, Å., and Xu, Y. (2010). The genetic dissection of quantitative traits in crops. *Electronic J. Biotechnol.* 13, 16–17. doi: 10.2225/vol13-issue5-fulltext-14

Shakoor, N., Lee, S., and Mockler, T. C. (2017). High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. *Curr. Opin. Plant Biol.* 38, 184–192. doi: 10.1016/j.pbi.2017.05.006

Shi, H., Mancuso, N., Spendlove, S., and Pasaniciu, B. (2017). Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101, 737–751. doi: 10.1016/j.ajhg.2017.09.022

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). *Nat. Genet.* 43, 109–116. doi: 10.1038/ng.740

Silva, P., Evers, B., Kieffaber, A., Wang, X., Brown, R., Gao, L., et al. (2022). Applied phenomics and genomics for improving barley yellow dwarf resistance in winter wheat. *G3* 12, jkac064. doi: 10.1093/g3journal/jkac064

Sokolkova, A., Burlyayeva, M., Valiannikova, T., Vishnyakova, M., Schafleitner, R., Lee, C.-R., et al. (2020). Genome-wide association study in accessions of the mini-core collection of mungbean (*Vigna radiata*) from the World Vegetable Gene Bank (Taiwan). *BMC Plant Biol.* 20, 1–9. doi: 10.1186/s12870-020-02579-x

Somta, P., Laosatit, K., Yuan, X., and Chen, X. (2022). Thirty years of mungbean genome research: Where do we stand and what have we learned? *Front. Plant Sci.* 13, 944721. doi: 10.3389/fpls.2022.944721

Song, J.-M., Guan, Z., Hu, J., Guo, C., Yang, Z., Wang, S., et al. (2020). Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants* 6, 34–45. doi: 10.1038/s41477-019-0577-7

Song, J. M., Liu, D. X., Xie, W. Z., Yang, Z., Guo, L., Liu, K., et al. (2021). BnPIR: *Brassica napus* pan-genome information resource for 1689 accessions. *Plant Biotechnol. J.* 19, 412. doi: 10.1111/pbi.13491

Sosiawan, H., Adi, S. H., and Yusuf, W. A. (2021). “Water-saving irrigation management for mung bean in acid soil,” in *IOP conference series: earth and environmental science* (Bristol, United Kingdom: IOP Publishing Ltd), vol. 648, No. 1, 012144.

Speed, D., and Balding, D. J. (2019). SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* 51, 277–284. doi: 10.1038/s41588-018-0279-5

Srungarapu, R., Mahendrakar, M. D., Mohammad, L. A., Chand, U., Jagarlamudi, V. R., Kondamudi, K. P., et al. (2022). Genome-wide association analysis reveals trait-linked markers for grain nutrient and agronomic traits in diverse set of chickpea germplasm. *Cells* 11. doi: 10.3390/cells11152457

Subramanian, R., and Narayana, M. (2023). Development of bruchid pest and mungbean yellow mosaic virus disease resistance lines in blackgram [*Vigna mungo* (L.) Hepper] through marker-assisted selection. *Physiol. Mol. Plant Pathol.* 127, 102105. doi: 10.1016/j.pmp.2023.102105

Sunitha, S., and Rock, C. D. (2020). CRISPR/Cas9-mediated targeted mutagenesis of TAS4 and MYBA7 loci in grapevine rootstock 101-14. *Transgenic Res.* 29, 355–367. doi: 10.1007/s11248-020-00196-w

Talakayala, A., Mekala, G. K., Reddy, M. K., Ankanagari, S., and Garladinne, M. (2022). Manipulating resistance to mungbean yellow mosaic virus in greengram (*Vigna*

- radiata L.): Through CRISPR/Cas9 mediated editing of the viral genome. *Front. Sustain. Food Syst.* 6, 911574. doi: 10.3389/fsufs.2022.911574
- Tapia, R., Abd-Elrahman, A., Osorio, L., Whitaker, V. M., and Lee, S. (2022). Combining canopy reflectance spectrometry and genome-wide prediction to increase response to selection for powdery mildew resistance in cultivated strawberry. *J. Exp. Bot.* 73, 5322–5335. doi: 10.1093/jxb/erac136
- Thabet, S. G., Sallam, A., Moursi, Y. S., Karam, M. A., Alqudah, A. M., and Wu, H. (2021). Genetic factors controlling nTiO₂ nanoparticles stress tolerance in barley (*Hordeum vulgare*) during seed germination and seedling development. *Funct. Plant Biol.* 48, 1288–1301. doi: 10.1071/FP21129
- Upadhyaya, H. D., Bajaj, D., Das, S., Kumar, V., Gowda, C., Sharma, S., et al. (2016). Genetic dissection of seed-iron and zinc concentrations in chickpea. *Sci. Rep.* 6, 1–12. doi: 10.1038/srep24050
- Vairam, N., Lavanya, S. A., and Vanniarajan, C. (2017). Screening for pod shattering in mutant population of mungbean (*Vigna radiata* (L.) Wilczek). *J. Appl. Natural Sci.* 9, 1787–1791. doi: 10.31018/jans.v9i3.1439
- Van, K., Kang, Y. J., Han, K.-S., Lee, Y.-H., Gwag, J.-G., Moon, J.-K., et al. (2013). Genome-wide SNP discovery in mungbean by Illumina HiSeq. *Theor. Appl. Genet.* 126, 2017–2027. doi: 10.1007/s00122-013-2114-9
- Van De Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., et al. (2019). A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell* 178, 1260–1272. doi: 10.1016/j.cell.2019.07.038
- Varoquaux, F., Blanvillain, R., Delseny, M., and Gallois, P. (2000). Less is better: new approaches for seedless fruit production. *Trends Biotechnol.* 18, 233–242. doi: 10.1016/S0167-7799(00)01448-7
- Vilhjálmsdóttir, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., et al. (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* 97, 576–592. doi: 10.1016/j.ajhg.2015.09.001
- Voicheck, Y., and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. *Nat. Genet.* 52, 534–540. doi: 10.1038/s41588-020-0612-7
- Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., et al. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature* 588, 277–283. doi: 10.1038/s41586-020-2961-x
- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet.* 17, e1009440. doi: 10.1371/journal.pgen.1009440
- Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, doi: 10.1038/s41467-017-01261-5
- Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat. Genet.* 53, 243–253. doi: 10.1038/s41588-020-00769-9
- Werme, J., van der Sluis, S., Posthuma, D., and De Leeuw, C. (2021). LAVA: An integrated framework for local genetic correlation analysis. *BioRxiv* 2012, 2031.424652. doi: 10.1101/2020.12.31.424652
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.18
- Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340
- Wolter, F., and Puchta, H. (2018). The CRISPR/Cas revolution reaches the RNA world: Cas13, a new Swiss Army knife for plant biologists. *Plant J.* 94, 767–775. doi: 10.1111/tj.2018.94.issue-5
- Wu, X., Feng, H., Wu, D., Yan, S., Zhang, P., Wang, W., et al. (2021a). Using high-throughput multiple optical phenotyping to decipher the genetic architecture of maize drought tolerance. *Genome Biol.* 22, 1–26. doi: 10.1186/s13059-021-02377-0
- Wu, J.-H., Ferguson, A. R., Murray, B. G., Jia, Y., Datson, P. M., and Zhang, J. (2012). Induced polyploidy dramatically increases the size and alters the shape of fruit in *Actinidia chinensis*. *Ann. Bot.* 109, 169–179. doi: 10.1093/aob/mcr256
- Wu, X., Islam, A. F., Limpot, N., Mackasmiel, L., Mierzwa, J., Cortés, A. J., et al. (2020b). Genome-wide SNP identification and association mapping for seed mineral concentration in mung bean (*Vigna radiata* L.). *Front. Genet.* 11, 656. doi: 10.3389/fgene.2020.00656
- Wu, X., Sun, T., Xu, W., Sun, Y., Wang, B., Wang, Y., et al. (2021b). Unraveling the genetic architecture of two complex, stomata-related drought-responsive traits by high-throughput physiological phenotyping and GWAS in cowpea (*Vigna unguiculata* L. Walp.). *Front. Genet.* 12, 743758. doi: 10.3389/fgene.2021.743758
- Wu, J., Wang, L., Fu, J., Chen, J., Wei, S., Zhang, S., et al. (2020a). Resequencing of 683 common bean genotypes identifies yield component trait associations across a north–south cline. *Nat. Genet.* 52, 118–125. doi: 10.1038/s41588-019-0546-0
- Wu, X.-T., Xiong, Z.-P., Chen, K.-X., Zhao, G.-R., Feng, K.-R., Li, X.-H., et al. (2022). Genome-wide identification and transcriptional expression profiles of PP2C in the barley (*Hordeum vulgare* L.) pan-genome. *Genes* 13, 834. doi: 10.3390/genes13050834
- Xiao, Q., Bai, X., Zhang, C., and He, Y. (2022). Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *J. Advanced Res.* 35, 215–230. doi: 10.1016/j.jare.2021.05.002
- Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., et al. (2020). Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. *Mol. Plant* 13, 187–214. doi: 10.1016/j.molp.2020.01.008
- Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88, 76–82. doi: 10.1016/j.ajhg.2010.11.011
- Youssef, H. M., Eggert, K., Koppolu, R., Alqudah, A. M., Poursarebani, N., Fazeli, A., et al. (2017). VRS2 regulates hormone-mediated inflorescence patterning in barley. *Nat. Genet.* 49, 157–161. doi: 10.1038/ng.3717
- Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38, 203–208. doi: 10.1038/ng1702
- Yuan, J., Kaur, D., Zhou, Z., Nagle, M., Kiddle, N. G., Doshi, N. A., et al. (2022). Robust high-throughput phenotyping with deep segmentation enabled by a web-based annotator. *Plant Phenomics*. doi: 10.34133/2022/9893639
- Yuan, J., Wang, X., Zhao, Y., Khan, N. U., Zhao, Z., Zhang, Y., et al. (2020). Genetic basis and identification of candidate genes for salt tolerance in rice by GWAS. *Sci. Rep.* 10, 1–9. doi: 10.1038/s41598-020-66604-7
- Zhang, Y., Lu, Q., Ye, Y., Huang, K., Liu, W., Wu, Y., et al. (2020). Local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *bioRxiv* 2005, 2008.084475. doi: 10.1101/2020.05.08.084475
- Zhang, Y., Wang, J., Du, J., Zhao, Y., Lu, X., Wen, W., et al. (2021). Dissecting the phenotypic components and genetic architecture of maize stem vascular bundles using high-throughput phenotypic analysis. *Plant Biotechnol. J.* 19, 35–50. doi: 10.1111/pbi.13437
- Zhang, Y., and Zhang, N. (2018). Imaging technologies for plant high-throughput phenotyping: a review. *Front. Agric. Sci. Eng.* 5, 406–419. doi: 10.15302/J-FASE-2018242
- Zhang, H.-X., Zhang, Y., and Yin, H. (2019). Genome editing with mRNA encoding ZFN, TALEN, and Cas9. *Mol. Ther.* 27, 735–746. doi: 10.1016/j.ymthe.2019.01.014
- Zhang, W., Zhao, Y., Yang, H., Liu, Y., Zhang, Y., Zhang, Z., et al. (2024). Comparison analysis of bioactive metabolites in soybean, pea, mung bean, and common beans: reveal the potential variations of their antioxidant property. *Food Chem.* 457, 140137. doi: 10.1016/j.foodchem.2024.140137
- Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., et al. (2018a). Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* 50, 278–284. doi: 10.1038/s41588-018-0041-z
- Zhao, X., Luo, L., Cao, Y., Liu, Y., Li, Y., Wu, W., et al. (2018b). Genome-wide association analysis and QTL mapping reveal the genetic control of cadmium accumulation in maize leaf. *BMC Genomics* 19, 1–13. doi: 10.1186/s12864-017-4395-x



OPEN ACCESS

EDITED BY

Huihui Li,
Chinese Academy of Agricultural Sciences,
China

REVIEWED BY

Wenxian Liu,
Lanzhou University, China
Gang Gao,
Chinese Academy of Agricultural
Sciences, China
Shugao Fan,
Ludong University, China

*CORRESPONDENCE

Guofeng Yang
✉ yanggf@qau.edu.cn
Zeng-Yu Wang
✉ zywang@qau.edu.cn
Kunlong Su
✉ sukl@qau.edu.cn

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

RECEIVED 09 July 2024

ACCEPTED 07 March 2025

PUBLISHED 01 April 2025

CITATION

Li Z, Yu Q, Ma Y, Miao F, Ma L, Li S, Zhang H,
Wang Z-Y, Yang G and Su K (2025) Screening
and functional characterization
of salt-tolerant NAC gene family
members in *Medicago sativa* L.
Front. Plant Sci. 16:1461735.
doi: 10.3389/fpls.2025.1461735

COPYRIGHT

© 2025 Li, Yu, Ma, Miao, Ma, Li, Zhang, Wang,
Yang and Su. 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.

Screening and functional characterization of salt-tolerant NAC gene family members in *Medicago sativa* L

Zhiguang Li^{1†}, Qianqian Yu^{1†}, Yue Ma¹, Fuhong Miao¹,
Lichao Ma^{1,2}, Shuo Li¹, Huajie Zhang^{3,4}, Zeng-Yu Wang^{1*},
Guofeng Yang^{1,3*} and Kunlong Su^{1*}

¹Key Laboratory of National Forestry and Grassland Administration on Grassland Resources and Ecology in the Yellow River Delta, College of Grassland Science, Qingdao Agricultural University, Qingdao, China, ²Academy of Dongying Efficient Agricultural Technology and Industry on Saline and Alkaline Land in Collaboration with Qingdao Agricultural University, Dongying, China, ³Weihai Animal Epidemic Disease Prevention and Control Center, Weihai, China, ⁴Weihai Academy of Agricultural Sciences, Weihai, China

Introduction: Alfalfa is the most widely cultivated high-quality perennial leguminous forage crop in the world. In China, saline-alkali land represents an important yet underutilized land resource. Cultivating salt-tolerant alfalfa varieties is crucial for the effective development and utilization of saline-alkali soils and for promoting the sustainable growth of grassland-livestock farming in these regions. The NAC (NAM, ATAF, and CUC) family of transcription factors plays a key role in regulating gene expression in response to various abiotic stresses, such as drought, salinity and extreme temperatures, thereby enhancing plant stress tolerance.

Methods: This study evaluated the structure and evolutionary relationship of the members of the NAC-like transcription factor family in alfalfa using bioinformatics. We identified 114 members of the NAC gene family in the Zhongmu No.1 genome and classified them into 13 subclasses ranging from I to XIII. The bioinformatics analysis showed that subfamily V might be related to the response to salt stress. Gene expression analysis was conducted using RNA-seq and qRT-PCR, and *MsNAC40* from subfamily V was chosen for further investigation into salt tolerance.

Results: *MsNAC40* gene had an open reading frame of 990 bp and encoded a protein containing 329 amino acids, with a molecular weight of 3.70 kDa and a conserved NAM structural domain. The protein was hydrophilic with no transmembrane structure. After treating both the *MsNAC40* overexpressing plants and the control group with 150 mmol/L NaCl for 15 days, physiological and biochemical measurements revealed that these plants had significantly greater height, net photosynthetic rate, stomatal conductance, and transpiration rate compared to the control group, while their conductivity was significantly lower. Additionally, the levels of abscisic acid in the roots and leaves, along with the activities of peroxidase, superoxide dismutase, and catalase in the leaves, were significantly higher in the overexpressing plants, whereas the malondialdehyde content was significantly lower. Moreover, the Na⁺ content

in the overexpressing plants was significantly reduced, while the K^+/Na^+ ratio was significantly increased compared to the control group.

Discussion: These results indicated that the *MsNAC40* gene improved the salt tolerance of *Pioneer Alfalfa SY4D*, but its potential mechanism of action still needs to be further explored.

KEYWORDS

alfalfa, NAC, salt tolerance, phylogenetic analysis, *MsNAC40*

1 Introduction

Salt-affected soils are a valuable land resource in China, covering approximately 10% of the country's total area. Effectively developing and utilizing these soils could significantly advance sustainable grassland agriculture. However, the salinity present in these soils presents major challenges to agricultural development (Shao et al., 2019). In this context, enhancing the salt tolerance of alfalfa is particularly important. Alfalfa, one of the most important perennial legume crops worldwide, is known for its high-quality forage production. In China, it plays a crucial role in the grassland industry and has significantly contributed to economic growth. As demand for alfalfa continues to rise, especially in regions with salt-affected soils, the need to boost alfalfa production has become increasingly urgent (Wan et al., 2023). Improving alfalfa's salt tolerance not only enables more effective use of these soils but also greatly aids in ecological restoration and land management in these areas.

Salt stress, a significant abiotic stressor, severely impacts plant growth and productivity by inducing osmotic stress and ion toxicity, leading to physiological drought and metabolic disruptions (Deinlein et al., 2014). To combat these challenges, plants have developed various adaptive mechanisms, including osmotic adjustment, maintenance of ion homeostasis, and management of oxidative stress (Zhang et al., 2022). Recent advancements in molecular biology and genetic engineering have furthered our understanding of the genetic mechanisms behind salt tolerance in alfalfa. These developments have facilitated the identification and manipulation of key genes, paving the way for enhanced stress resilience in this vital crop.

NAC transcription factors are one of the largest families of transcriptional regulators widely found in plants and have been shown to be involved in various plant growth and developmental processes and abiotic stress responses (Diao et al., 2020). The acronym NAC is derived from the names of three genes containing specific structural domains: NAM (no apical meristem), ATAF1/ATAF2 (Arabidopsis transcription ACTivation factor 1/2), and CUC2 (cup-shaped cotyledon 2). Several NAC genes have been identified in various plants, including *Arabidopsis thaliana* (117), rice (151), grapevine (79), citrus (26), grape (26), poplar (163), soybean (152), and tobacco (152) (Hu et al., 2010; Le et al., 2011; Nuruzzaman et al., 2010; Rushton et al., 2008). Previous studies have revealed that members of the NAC

transcription factor family are extensively involved in the regulation of growth and developmental processes in different plants, including seed development, embryo development, stem tip meristem formation, stem fibre development, leaf senescence, and cell division (Duval et al., 2002; Guo et al., 2005; Kim et al., 2007, 2006; Ko et al., 2007; Sperotto et al., 2009). In addition, it has been shown that the NAC transcription factors can regulate plant responses to various biotic and abiotic stresses in different plants. For example, *ANAC019*, *ANAC055*, and *ANAC072* positively regulate drought tolerance, salt tolerance and abscisic acid content in *Arabidopsis* (Tran et al., 2004). Several other NAC-like transcription factors are also related to stress tolerance in *Arabidopsis*. For example, *ANAC083*, *ANAC041*, *ANAC054*, and *ANAC084* positively regulate seed germination under salt stress (Balazadeh et al., 2010), while *NAC1* positively regulates growth hormone and root development (Guo et al., 2005), and *ANAC019*, *ANAC042*, and *ANAC102* positively regulate cold stress, heat stress, and waterlogging, respectively (Christianson et al., 2009; Jensen et al., 2010; Shahnejat-Bushehri et al., 2017; Yuan et al., 2019). NAC-like TFs typically refer to transcription factors that share similar domains or functions with the NAC transcription factor family. While they possess similar NAM domains, they may differ slightly in evolution or function. NAC-like transcription factors have also been studied more extensively in maize, rice, and soybeans. In soybeans, *GmNAC11*, *GmSIN1*, and *GmNAC20* can improve the salt tolerance, while *GmNAC20* can also improve the cold tolerance of the plants (Hao et al., 2011; Li et al., 2019). In rice, the regulatory effects of NAC-like transcription factors, such as *OsNAC4*, *OsNAC5*, *OsNAC6*, and *OsNAC10*, on abiotic stresses increase stress tolerance (Hu et al., 2008; Jeong et al., 2010; Sperotto et al., 2009; Zheng et al., 2009). *ZmNAC1* positively regulates low-temperature, high-salt, drought, and ABA stresses in maize (Lu et al., 2012).

In this experiment, we analysed the structure and evolutionary relationship of the members of the NAC-like transcription factor family in alfalfa using bioinformatics by screening and evaluating the relative expression of 15 candidate genes that were responsive to salt stress based on the transcriptome of the salinity-tolerant alfalfa. We also investigated the relative expression of these genes in various tissues of alfalfa at various time points under salt treatment. We transformed *MsNAC40* into alfalfa using an overexpression vector to obtain transgenic material overexpressing the *MsNAC40* gene. The positive transgenic lines overexpressing the *MsNAC40* gene

were subjected to phenotypic, physiological, biochemical, and metabolic analyses to preliminarily investigate the function of the gene in salinity tolerance. The study identified candidate genes for salt tolerance in alfalfa and provides a theoretical basis for the selection and breeding of alfalfa varieties for salt tolerance.

2 Materials and methods

2.1 Identification of the NAC family members in alfalfa

The Zhongmu No.1 whole protein sequence, genome open reading frame, and genome annotation file were obtained from the alfalfa Zhongmu No.1 genome website (<https://modms.lzu.edu.cn/>) (Fang et al., 2024). The HiddenMarkovModel profile (E-value $>e^{-10}$) of NAM (PF02365) (Wang et al., 2010) was downloaded from the Pfam website (<https://pfam.xfam.org/>) to identify the members of the NAC gene family in the alfalfa genome, and to collect and analyze the gene names, gene IDs, and gene annotations associated with each identified NAC transcription factor. Moreover, the gene name, gene ID, number of amino acids encoded, molecular weight, isoelectric point, instability coefficient, fat coefficient, and total average hydrophilicity physicochemical indexes of each NAC gene family transcription factor were also compiled and analysed. To ensure the accuracy of the selected genes, the predicted NAC protein sequences will be submitted to InterProScan (<http://www.ebi.ac.uk/interpro/serach/sequence-serach>), CDD (<http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>), Pfam, and SMART (<http://smart.embl-heidelberg.de/>) for sequence calibration and verification of conserved domains. Subsequently, the isoelectric points and molecular weights of the NAC family protein members will be analyzed using the ExPASy (http://web.expasy.org/compute_pi/) website.

2.2 Conserved motifs and gene structure analysis of the alfalfa NAC gene family

The conserved domains of the alfalfa NAC gene family were searched in the NCBI database (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) to further analyse the *Medicago sativa* NACs (*MsNACs*). A phylogenetic tree of alfalfa NAC gene family members was constructed via MEGA11.0 using the Neighbor-Joining (NJ) method with the bootstrap value set to 1000.

2.3 Chromosomal localisation and covariance analysis of the alfalfa NAC gene members

The chromosomal positions of alfalfa NAC gene members were screened based on the genome annotation information of Zhongmu No.1, and the TBtools software was used to map the chromosomal localisation of alfalfa NAC genes. *Medicago polymorpha*,

Arabidopsis thaliana, and *Medicago sativa* genomes were subjected to covariance analysis, and the interspecies and intraspecies similar genes were analysed functionally.

2.4 Analysis of the cis-acting elements of the alfalfa NAC gene members

The cis-acting elements located 2000 bp upstream of the gene region of the alfalfa NAC gene family members were analysed based on the Zhongmu No.1 genome obtained from the PlantCARE website (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html/>) (Lescot et al., 2002). The cis-acting elements were analysed using the TBtools (Chen et al., 2020), and the different types of cis-acting elements were visualized using a heat map generated using TBtools.

2.5 Screening and expression analysis of *MsNAC* transcription factors under salt and alkali stress in alfalfa

To identify *MsNAC* transcription factors with different expression levels and screen for *MsNAC* genes with large expression differences between salt stress and alkali stress, we used transcriptome data of alfalfa subjected to salt and alkali stresses, as reported by the College of Grassland Science of Qingdao Agricultural University. The transcriptome data were generated from 4-week-old alfalfa seedlings subjected to salt, alkali, and salt-alkali mixed treatments. The treatments were divided into the following seven groups: Group A (control group), Group B (100 mmol/L NaCl solution), Group C (100 mmol/L NaHCO₃ solution), Group D (90 mmol/L NaCl+10 mmol/L NaHCO₃ solution), Group E (80 mmol/L NaCl+20 mmol/L NaHCO₃ solution), Group F (70 mmol/L NaCl +30 mmol/L NaHCO₃ solution), and Group G (60 mmol/L NaCl+40 mmol/L NaHCO₃ solution). Samples were taken at days 1 and 6 for transcriptome sequencing. A heat map showing the expression of *MsNAC* genes was generated using TBtools, and Table 1 presents the list of primers and their sequences.

2.6 Analysis of expression patterns of candidate *MsNAC* genes under salt treatment in root and leaf tissues

RNA was extracted using the Vazyme kit from the 2nd and 3rd leaves of the stem apical part and the 3 cm region of the root tips of 4-week-old hydroponic seedlings of Zhongmu No.1 sampled at 0h, 12h, 24h, and 48h after 50 mmol/L NaCl, 100 mmol/L NaCl, and 150 mmol/L NaCl salt treatment, as well as those subjected to double distilled water (ddH₂O) incubation (control). The treatment procedures are detailed in Table 2. The RNA samples were reverse transcribed into cDNA for real-time fluorescence quantitative PCR analysis. The reaction conditions were as follows: *MsActin* was used as the internal control gene, and *MsNAC40* was the target gene. The relative expression levels of the genes in each group were determined using the Ct method, and the expression levels of the genes were

TABLE 1 List of primers and their sequences.

Primername	Primersequence
Actin-F	ACTGGAATGGTGAAGGCTGG
Actin-R	TGACAATACCGTGCTCAATGG
qMsNAC4-F	TCATTACTTTTATTTGC
qMsNAC4-R	ATCTCTTTATCTTTTCCA
qMsNAC30-F	TTGGAAAGCAACTGGAAA
qMsNAC30-R	CACGAGGGCTAAAGAAAT
qMsNAC29-F	AAGTTACCACCCTGTTTT
qMsNAC29-R	GTCTCCTCCCGTTTTTG
qMsNAC40-F	TTCCAGAGAGAGATCCTC
qMsNAC40-R	CTCACCAAAATTCGCCTT
qMsNAC39-F	CACCTGGTTTCAGATTCT
qMsNAC39-R	CCCTCAACTTTTCTTTTT
qMsNAC50-F	AGCTTGATGTTATTCCAG
qMsNAC50-R	AATCTTTCTCTCTTTTCC
qMsNAC51-F	GGACACAAAATAGAATGA
qMsNAC51-R	CTAGAGGAAGAAGCAGAA
qMsNAC52-F	TGGGTTTGCTTCTCTCC
qMsNAC52-R	GATTTATTTTCTCTCACT
qMsNAC85-F	TGGCAAGACCAAGTTTTT
qMsNAC85-R	GGTATGATGCTAGGATGA
qMsNAC79-F	ATTCTCCTCAGCTCTGTG
qMsNAC79-R	CTTTCTGCCTGCTCTCTT
qMsNAC70-F	TAAGGTCTTCTCTTTCCC
qMsNAC70-R	AACCAGTTGCTTTCCAGT
qMsNAC77-F	GATTGCCTCCTGGTTTTT
qMsNAC77-R	TGGCTTCCTTGCTGCTGA
qMsNAC78-F	ACAACAACAAGGAGAAAAG
qMsNAC78-R	AGGTAATGAAATGGAAAT
qMsNAC108-F	TGGACACAGCCAAGACAG
qMsNAC108-R	GGGACACCAACAACAGCA
qMsNAC113-F	TGCTTCACACTTTTTCCA
qMsNAC113-R	GCTTTTCTCCTCACTCTCC

calculated using Excel. The variability of the treatment groups was calculated using IBM SPSS Statistics25 software.

2.7 Cloning and protein structure analysis of the *MsNAC40* gene

The *MsNAC40* protein sequences were uploaded to ExPASy protparam (<https://www.expasy.org/resources/protparam>), ExPASy

TABLE 2 List of Salt stress treatment methods and sampling sites.

Number	Salt treatment method	Sampling location
CK1	ddH2O	Root tip 2 to 3 cm
A1	50 mmol/L NaCl	Root tip 2 to 3 cm
B1	100 mmol/L NaCl	Root tip 2 to 3 cm
C1	150 mmol/L NaCl	Root tip 2 to 3 cm
CK2	ddH2O	The second and third leaves at the top of the stem
A2	50 mmol/L NaCl	The second and third leaves at the top of the stem
B2	100 mmol/L NaCl	The second and third leaves at the top of the stem
C2	150 mmol/L NaCl	The second and third leaves at the top of the stem

protscale (<https://www.expasy.org/resources/protscale>), and NetPhos3.1 (<https://services.healthtech.dtu.dk/services/NetPhos-3.1/>) platforms to analyse the primary structure, hydrophilicity and the phosphorylation sites of the *MsNAC40* proteins, respectively. Bio Lib (<https://dtu.biolib.com/DeepTMHMM>) was used for phosphorylation site mapping of the *MsNAC40* proteins, while TMHMM2.0 (<https://services.healthtech.dtu.dk/services/TMHMM-2.0/>) was used to predict the transmembrane helical structure of the *MsNAC40* proteins. Moreover, SWISS-MODEL (<https://swissmodel.expasy.org/>) was used to predict the tertiary structures of the *MsNAC40* proteins.

2.8 Tissue-specific expression of *MsNAC40*

Alfalfa SY4D was cultured to the early flowering stage, after which RNA was extracted from the root, stem, leaf, flower, and branch tissues of the plants. The RNA samples were reverse transcribed into cDNA, and real-time fluorescence-based quantitative PCR experiments were conducted, with each sample being repeated three times. The sample data were normalized based on the internal control *MsAction*, and the relative expression levels

TABLE 3 List of primers.

Primer name	Primer sequence
MsNAC40-F	ATGGGAGTTCAGAGAGAGATCCTC
MsNAC40-R	TTAATGACCCGAATACCCAAACC
M13F	TGTAAACGACGGCCAGT
M13R	CAGGAAACAGCTATGACC
Actin-F	ACTGGAATGGTGAAGGCTGG
Actin-R	TGACAATACCGTGCTCAATGG
qMsNAC40-F	TTCCAGAGAGAGATCCTC
qMsNAC40-R	CTCACCAAAATTCGCCTT

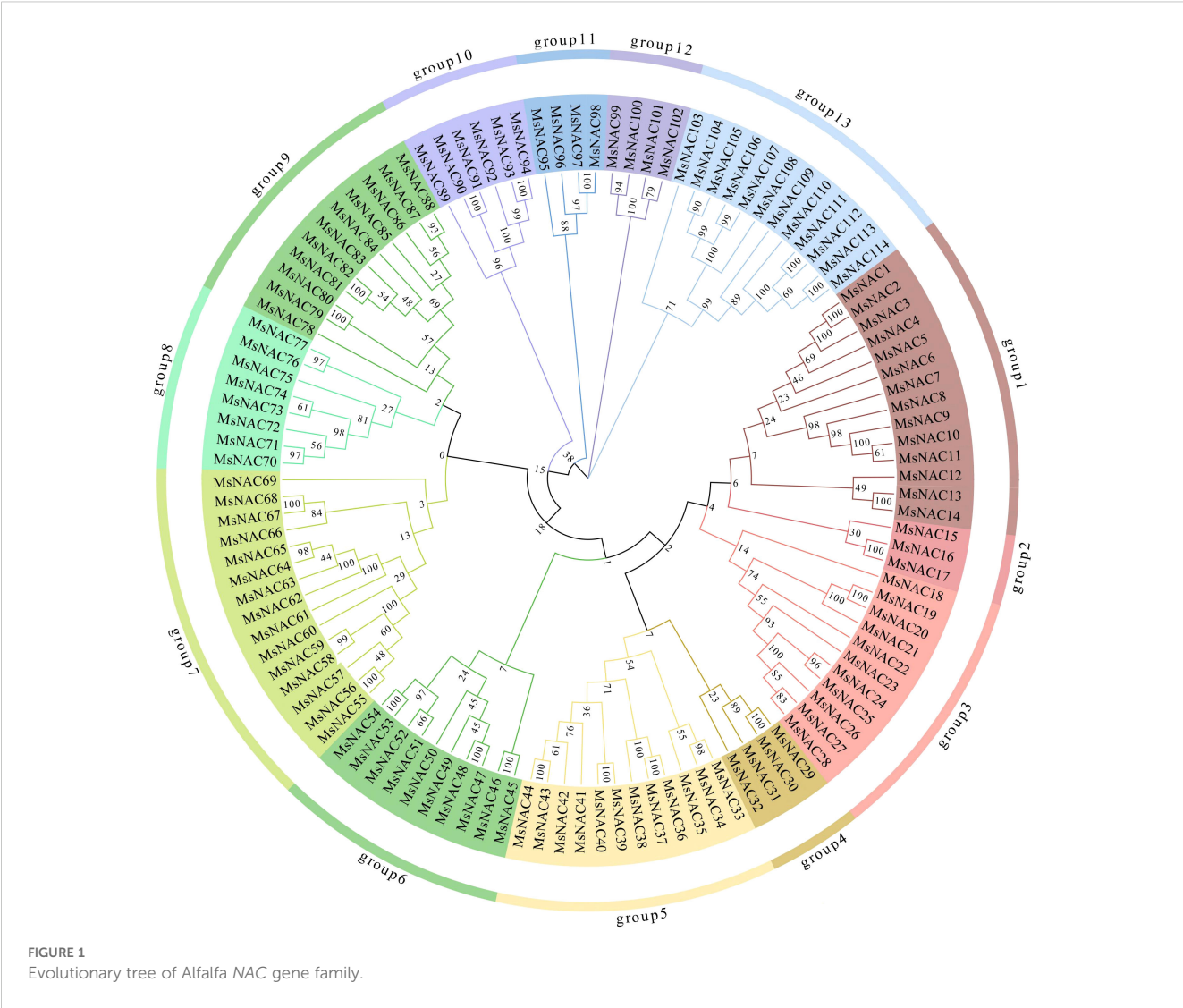


FIGURE 1
Evolutionary tree of Alfalfa NAC gene family.

of the genes in each group were determined using the Ct method. The expression levels of the genes were calculated using Excel, and the IBM SPSS Statistics25 software was used to calculate the variability of each treatment group.

2.9 Identification of positive seedlings and expression analysis of positive transgenic plants

The 3301MsNAC40-F/R primers were designed with NcoI and PmlI enzymatic cleavage sites (Table 3) to obtain the target gene fragment with homologous arms for ligation into the linear vector pCAMIBA3301 digested with NcoI and PmlI restriction enzymes. Colony PCR of the transformed E. coli cells detected no error, and the E. coli cells were transferred into Agrobacterium EHA105.

The vector was constructed and transformed into Agrobacterium EHA105 cells, which were then used for the leaf disc transformation of young leaves of 4-weeks-old alfalfa SY4D. The transformed leaves were

co-cultured in SH3a media for 20 h in the dark and then transferred into the selection medium (attachment). The formed calluses were cultured in the dark for 2 to 3 months and were transferred to the MSBK medium for about one week under the photoperiod cycle of 16 h light/8 h dark for 30 to 45 d. The green-sprouting calluses were transferred to SH9a medium (attachment) until the tissues regenerated into plantlets. The regenerated plantlets were then grown in the glasshouse.

Thereafter, DNA was extracted from the young leaves of the regenerated plants for PCR analysis using primers 3301JY-F/R and 31JY-F/R to confirm the presence of the transformed gene. The bands matching the length of the target fragment were sequenced and compared with the target sequence to their similarity. Moreover, RNA was extracted from the leaf tissues of the transgenic plants overexpressing the target gene and wild-type alfalfa SY4D and reverse transcribed into cDNA. qPCR was performed to determine the expression level of the target gene in the transgenic alfalfa plants using qMsNAC40-F/R and Actin-F/R primers, and three overexpression plants with higher expression levels were selected.

2.10 Salt tolerance phenotype analysis of the overexpression alfalfa plants

Overexpression alfalfa SY4D plants with uniform growth were selected, and 9-cm cuttings were planted in nutrient soil in three pots. The pots were kept in the greenhouse for 30 d after daily treatment with 100ml of half-strength (1/2) Hoagland's nutrient solution containing 150 mmol/L NaCl for 15 days. Thereafter, the absolute height from the ground to the highest point of the plant was measured in triplicate using a straightedge, and the three measurements were averaged. Fresh weight was measured by cutting the above-ground portion of the plant flush from the ground with scissors. The measurement was conducted in triplicate, and the three measurements were averaged.

2.11 Analysis of the physiological indicators of salt tolerance in alfalfa plants overexpressing the target gene

The plants were treated as described in section 2.10, and the apical 2nd and 3rd leaves were sampled at 10:30–11:00 a.m. to determine photosynthetic indexes, including net photosynthetic rate, stomatal conductance, and transpiration rate. Ten leaves were sampled from each treatment, and the average value was determined. Furthermore, 0.3g of fresh leaves were weighed and cut into small sections (1.5 cm) to determine the initial conductivity E1. The leaves were then incubated in boiling water for 20 min to cool down to determine the second conductivity E2. The measurements were repeated 3 times, and the relative conductivity was then calculated as $(E2-E1)/E2$.

2.12 Analysis of the biochemical indicators of salt tolerance and the contents of K^+ and Na^+ in root and leaf tissues of alfalfa plants overexpressing the target gene

The plants were treated as described in section 2.10. Appropriate amounts of leaves were sampled for measuring the proline, malondialdehyde, peroxidase, superoxide dismutase, and catalase contents using the Solepol activity test kit. After 12h of treatment with 100ml of 1/2 Hoagland nutrient solution containing 150 mmol/L NaCl, the leaves and roots were collected for measuring the abscisic acid contents using the Solepol Absciscic Acid Activity Test Kit.

The plants were treated as described in section 2.10, and the root tip tissues were sampled and oven-dried for 1h at 105°C, followed by 24h at 80°C. Thereafter, the samples were ground to powder and weighed (indicate the amount weighed here) for a 2h digestion using 1mL of HNO_3 and 1 mL of H_2O_2 . The supernatant was collected and left standing overnight. The K^+ and Na^+ standard solutions were diluted to 0, 5, 10, 20, 30, 40, and 50 $\mu\text{g/mL}$ with ultrapure water, and the standard curve was plotted. The samples were diluted 10 times with ultrapure water, and the concentrations of K^+ and Na^+ were determined using an M425 flame spectrophotometer. The final

contents of K^+ and Na^+ in alfalfa roots were calculated as $[C \text{ (measured concentration)} \times V \text{ (volume of measured liquid)} \times K \text{ (fractionation times)}]/m \text{ (mass of dried sample)}$.

3 Results

3.1 Identification of NAC family members in alfalfa

We screened 114 NAC genes from the Zhongmu No.1 genome, as shown in Table 1. The statistical results showed that the number of amino acids, the molecular weight, the isoelectric point, the instability coefficient, and fat coefficients of the alfalfa NAC family members ranged from 64 to 1094aa, 7.24 to 124.7 kDa, 4.09 to 9.84, 21.68 to 62.88, and 51.61 to 89.88, respectively (Supplementary Table S1). According to the prediction results, all NAC family members were hydrophilic proteins. Moreover, we analysed the results of the subfamily classification of *Medicago truncatula*, *A. thaliana*, and Soybean and found that the alfalfa NAC family classifies into 13 subfamilies, named subfamilies I to XIII (Le et al., 2011; Ling et al., 2017a, 2017; Ooka et al., 2003). The number of gene members in each I–XIII subfamilies was 14, 3, 11, 4, 12, 10, 15, 8, 11, 6, 4, 4, and 12, in that order, and the genes were named according to their order in the evolutionary tree, from *MsNAC1* to *MsNAC114* genes Figure 1.

3.2 Conserved motifs and gene structure analysis of the alfalfa NAC gene family

As shown in Figure 2, most of the conserved domains of the NAC genes matched the NAM sequences. The amino acid deletions were classified into three types: deletions at the beginning of *MsNAC50*, *MsNAC53*, *MsNAC54*, and *MsNAC86*, deletions at the end of *MsNAC22*, *MsNAC60*, *MsNAC75*, *MsNAC76*, and *MsNAC99*, and deletions at the middle of *MsNAC26*, *MsNAC64*, and *MsNAC65*.

The 114 NAC-like genes were subjected to motif analysis, and eight high-confidence motifs, named motif1–motif8, were selected for further analysis. It was found that most of the NAC gene family members contained these eight motifs, and the most abundant motif was motif 1, followed by motif 5, motif 8, motif 3, motif 4, motif 6, motif 2, and motif 7. The gene structure of the alfalfa NAC family members was analysed, and as shown in Figure 3, all alfalfa NAC family genes contained introns, and more genes contained more than five introns. Subfamilies V, VIII, and XII had relatively simple gene structures and high structural similarity within the subgroups.

3.3 Chromosomal localisation and covariance analysis of the alfalfa NAC gene members

As shown in Figure 4, the number of genes on chromosomes 1 through 8 were 15, 14, 14, 16, 14, 8, 14, and 17, respectively. Three pairs of tandem duplications (25/28, 43/44 and 111/114) and 21



The genome sequences of the widely studied model plant, *A. thaliana*, and the model legume plant, *Medicago sativa*, were used as

frontiersin.org

No.1 and Xinjiang Large-Leaf alfalfa. The observed gene distribution aligns with patterns seen in other polyploid species, such as cotton (Wu et al., 2013), and may reflect evolutionary pressures and functional adaptations of polyploid plants under abiotic stress conditions. Finally, based on the expression levels of candidate genes, *MsNAC40* from subfamily V was selected as the focus for further research.

3.4 Analysis of the cis-acting elements of the alfalfa NAC gene members

We analysed the sequence of the 2000 bp upstream of the promoter region of the alfalfa NAC genes. As shown in Figure 6, the cis-acting elements were mainly classified into three categories: (1) the plant growth and development category, which contained photosynthesis-related elements, such as G-box, Box4, GT1-motif, etc; (2) the plant hormone response category containing elements related to abscisic acid, gibberellin and other hormone responses, such as ABRE, P-box, etc; (3) the abiotic and biotic stress category containing anaerobic induction and other abiotic response-related elements, such as ARE, DRE, MBS, TC-richrepeats, etc. Through cluster analysis, we found that the NAC gene members in subfamily V contained more ABRE and ARE elements, suggesting that they may regulate the responses to abscisic acid and anaerobic stress. Moreover, subfamily V members contained many drought- and salt-stress-responsive elements, while those of subfamilies VI, VII, VIII, and XIII contained more LTR elements. Subfamily XIII members contained more MBS elements.

3.5 Screening and expression analysis of *MsNAC* transcription factors under salt and alkali stress in alfalfa

The transcriptomic data, which was obtained from the College of Grassland Agriculture at Qingdao Agricultural University, identified 74 NAC transcription factors with significantly different expression levels. As illustrated in Figure 7, sequence alignment was used to map these 74 NAC transcription factors to the categorized alfalfa NAC gene family, resulting in 74 of the 114 genes having corresponding data. The data were row normalised, after which 12 genes (4, 29, 30, 39, 40, 76, 77, 78, 79, 85, 108, and 113) with large differences in their expression levels under salt stress and three genes (50, 51, and 52) with large differences in their expression levels under alkali stress and mixed saline and alkaline stress were screened.

3.6 Analysis of expression patterns of candidate *MsNAC* genes under salt treatment in root and leaf tissues

We further analysed the expression levels of 15 candidate genes in root and leaf tissues under different concentrations of salt treatment. The results showed that the expression levels of seven genes (*MsNAC39*,

MsNAC40, *MsNAC76*, *MsNAC77*, *MsNAC78*, *MsNAC85*, and *MsNAC108*) in the root and leaf tissues, expression levels of *MsNAC51* in the leaf tissues and the expression levels of *MsNAC79* in the root tissues increased under salt treatment (Figure 8). Moreover, the expression levels of *MsNAC40* and *MsNAC78* in the root and leaf tissues, the expression of *MsNAC77* in the leaf tissues, and the expression of *MsNAC79* in the root tissues increased with the increasing salt concentration. The highest fold change of the relative expression of *MsNAC40* in the root and leaf tissues was more than 8-fold. Thus, based on these results and those of transcriptomic data and cis-acting elements analysis, *MsNAC40* was selected for further analysis.

3.7 Cloning and protein structure analysis of the *MsNAC40* gene

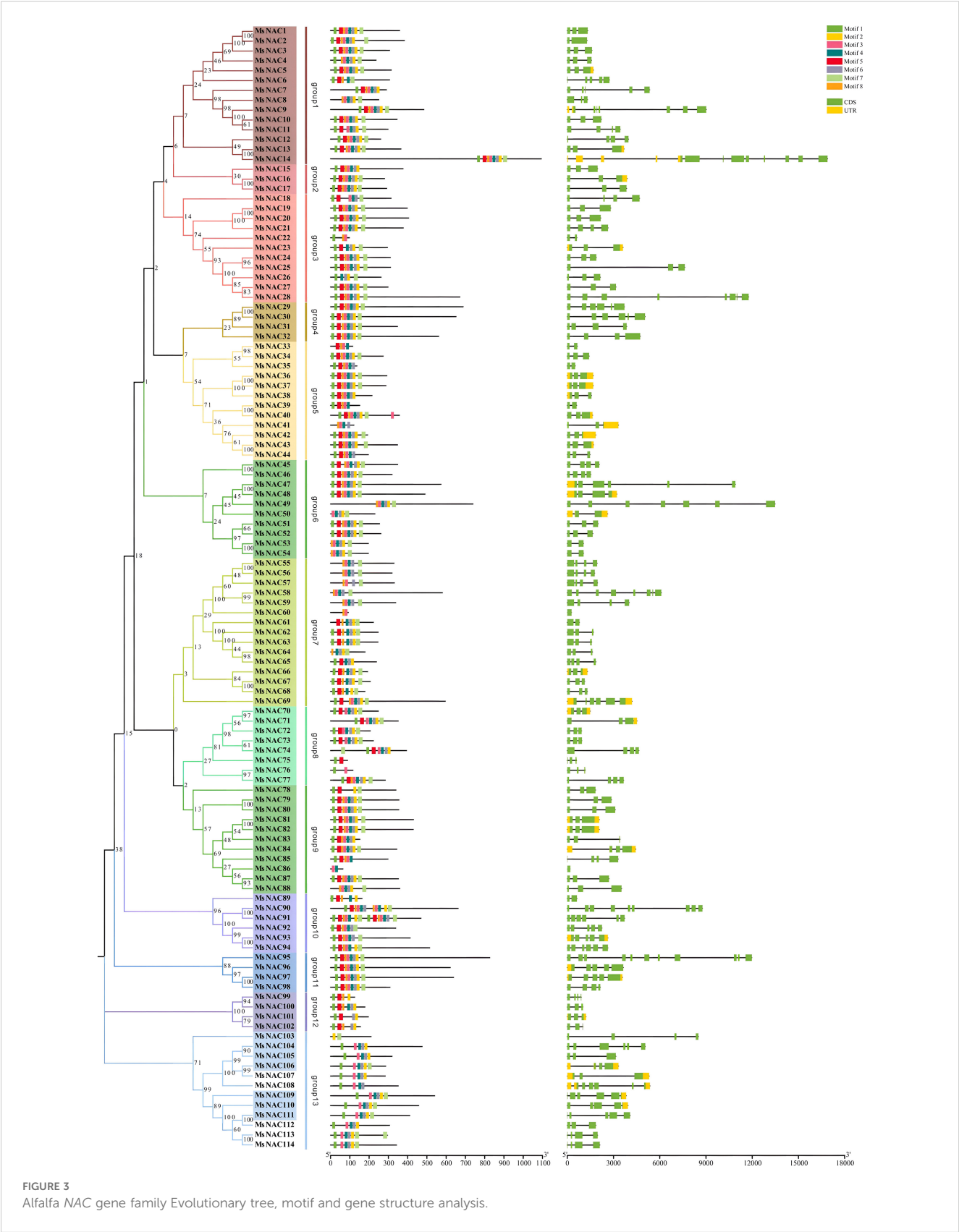
As shown in Figure 9, the *MsNAC40* gene was cloned from alfalfa SY4D, sequenced, and compared with the protein sequence encoded by Zhongmu No.1 *MsNAC40* (MsG0880044354.01.T01). The results showed that the cloned *MsNAC40* gene had 99% similarity with the Zhongmu No.1 *MsNAC40*. The length of the complete open reading frame of *MsNAC40* was 990bp, encoding 329 amino acids, and its protein structural domain was NAM (PF02365), belonging to the NAC-like gene family. *MsNAC40* was classified under the subfamily V in the alfalfa NAC family evolutionary tree.

Furthermore, the *MsNAC40* protein was 96.96% homologous to *MtNAC3* of *T. terrestris* alfalfa. The SWISS-MODEL results showed that the rice *NAC1* protein model 3ulx.1.B was the homologous model of *MsNAC40* protein, with a GMQE value of 0.69 and a similarity of 67.86%. As shown in Figures 9B, C, the NAM conserved domain overlapped the area of high prediction confidence, indicating that the functionally conserved NAM domains of the *MsNAC40* protein have high similarity with the rice *NAC1* protein model 3ulx. The physicochemical properties of *MsNAC40* protein were analysed using ExPASy (Figures 9D, E). We found that the number of encoding amino acids was 329, and the protein had a molecular weight of 3.70 KDa, an isoelectric point of 6.19, a total number of negatively charged amino acid residues (Asp+Glu) of 39, and a total number of positively charged amino acid residues (Arg+Lys) of 36. The hydrophilicity of the *MsNAC40* protein was analysed, and it was found that the 78th amino acid residue had the best hydrophilicity (-2.800), while the 45th amino acid had the best hydrophobicity (1.778). Additionally, there were more hydrophilic amino acid residues than hydrophobic ones, and the *MsNAC40* protein had no transmembrane structure.

Protein phosphorylation sites on the *MsNAC40* protein were predicted via the NetPhos website. As shown in Figure 9E, there were 40 phosphorylation sites on the *MsNAC40* protein, of which the serine, threonine, and tyrosine phosphorylation sites were 27, 8, and 5, respectively.

3.8 Tissue-specific expression of *MsNAC40*

We analysed the expression level of *MsNAC40* in the leaf, root, flower, stem, and branch tissues of 4-week-old alfalfa plants. As shown



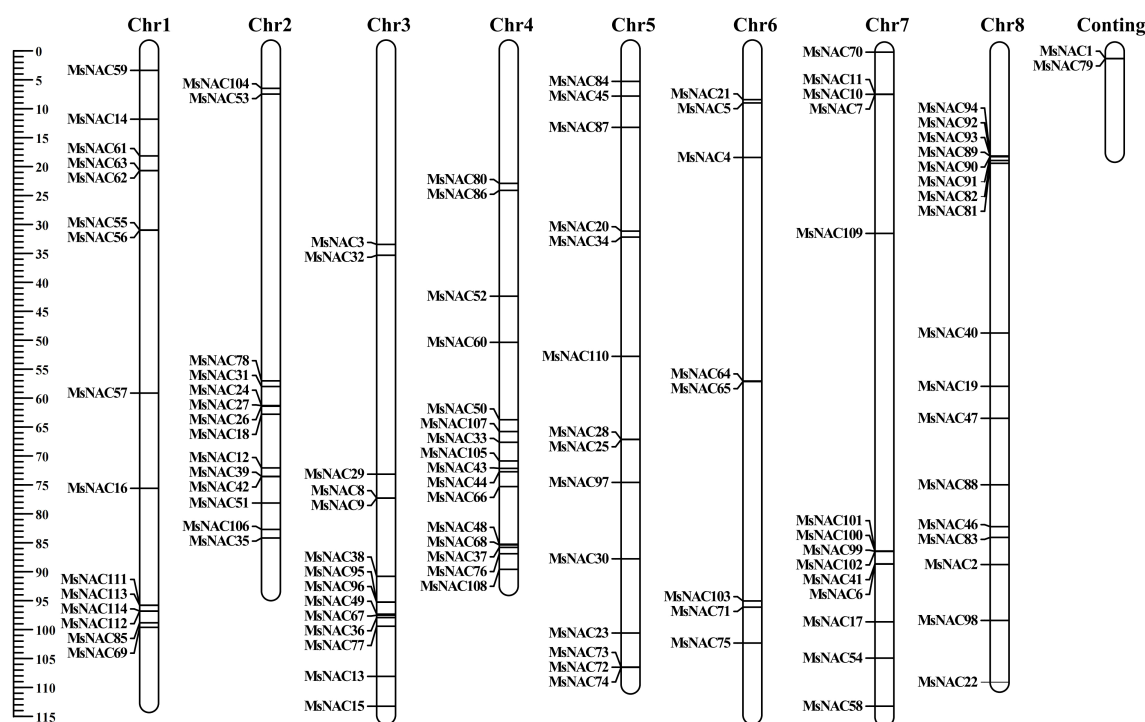


FIGURE 4
Chromosome mapping of Alfalfa NAC gene family.

in Figure 10, the relative expression levels of the target gene in the different tissues were ordered as roots>stems> branches> leaves>flowers, indicating the relative expression of the gene in root tissues was significantly higher but significantly lower in flower tissues than in the other tissues. This suggests that *MsNAC40* may play a primary function in plant roots and a secondary function in stem, branch and leaf tissues.

3.9 Identification of positive seedlings and expression analysis of positive transgenic plants

As depicted in Supplementary Figure S3, through the operation of Agrobacterium-mediated leaf disc transformation, we found that 35 *MsNAC40*-overexpressing alfalfa seedlings were positive for the target gene (Supplementary Figure S3A). The first batch of alfalfa seedlings used for the identification of positive plants contained 1 to 7 lines, and sequencing results showed that the amplified gene from the seven lines was consistent with the target sequence (Supplementary Figure S3B), similar to those amplified from lines 8 to 35 (Supplementary Figure S3C).

The qPCR experiments were also performed on the 35 overexpression lines, and the three lines with the highest expression were selected for subsequent analysis. As shown in Figure 11, lines d5, d6 and d8 had higher expression levels and were named L5, L6 and L8.

3.10 Salt tolerance phenotyping of the alfalfa plants overexpressing the target gene

As shown in Figure 12, there was no difference in the plant height and fresh weight between the control alfalfa plant SY4D and the transgenic lines under control conditions. However, after 15d of salt treatment, the fresh weight and plant height of line 8 was significantly higher than that of the control, and the plant height of lines 5 and 6 was significantly higher than that of the control. This indicated that *MsNAC40*-overexpressing alfalfa plants grew better under the 150 mmol/L NaCl treatment.

3.11 Analysis of the physiological indicators of salt tolerance in alfalfa plants overexpressing the target gene

As shown in Figure 13, the photosynthetic indexes of the three lines were significantly lower, but their conductivity was significantly higher than that of the controls after 15d of salt treatment. Based on the comprehensive analysis of the photosynthesis indexes and conductivity, the conductivity of wild-type alfalfa SY4D was significantly higher than that of the overexpression lines after salt treatment. The net photosynthetic rate, stomatal conductance and transpiration rate of the wild-type alfalfa SY4D were significantly lower

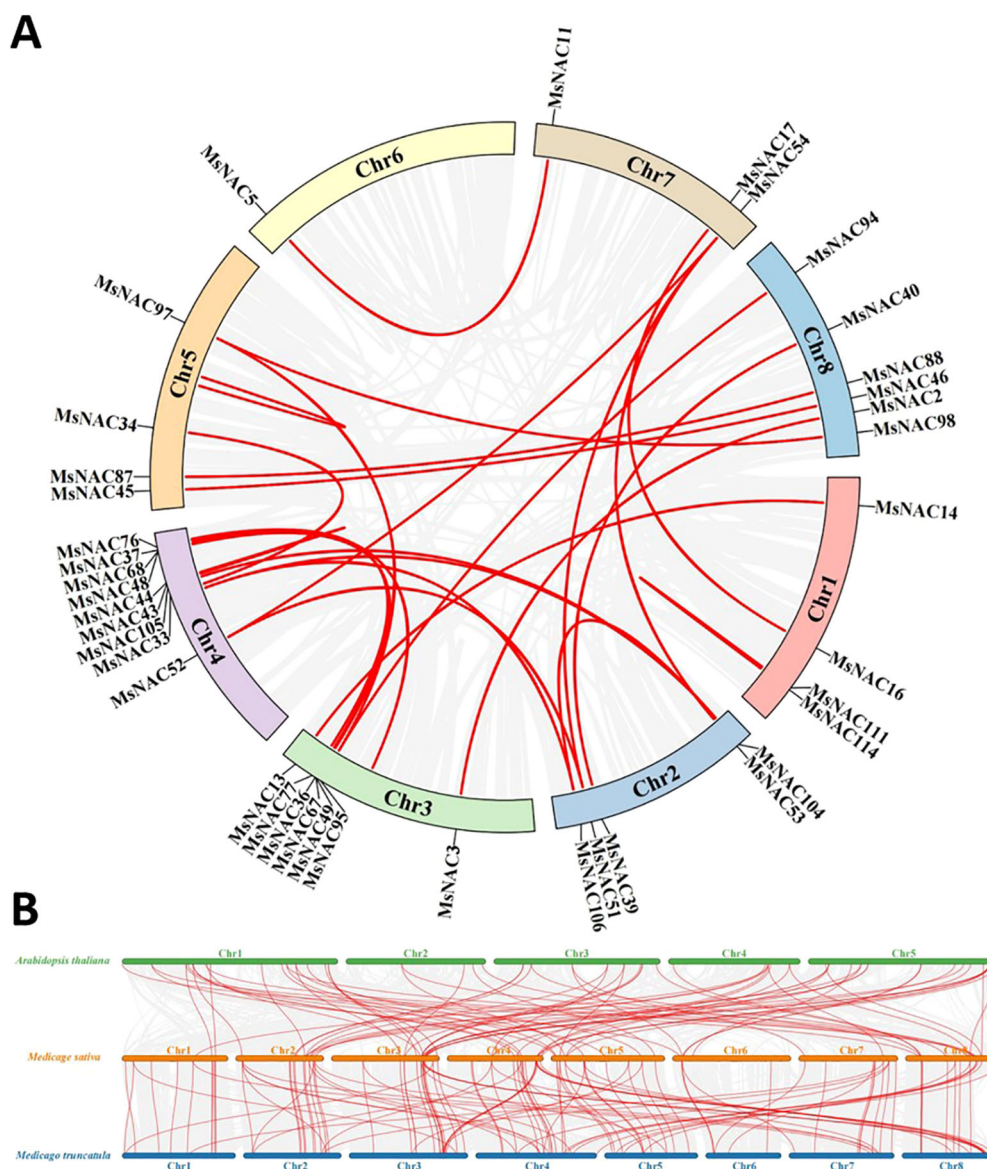


FIGURE 5

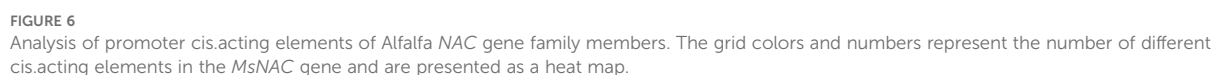
Collinearity analysis of Alfalfa NAC gene family. Red is collinear gene in gene family; (A) shows the collinearity of the NAC gene family, and (B) shows the collinearity analysis of Alfalfa with Arabidopsis and Tribulus. The gray line is the collinear block of Arabidopsis Tribulus and Alfalfa, and the red line is the NAC homologous gene pair.

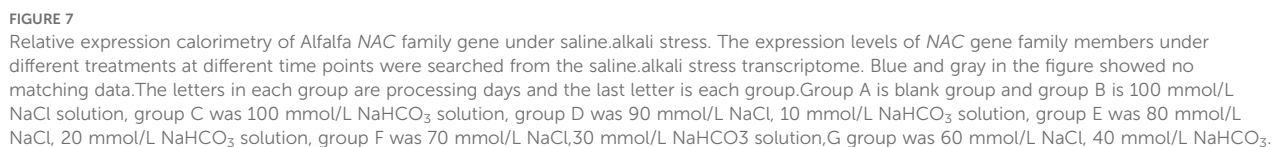
than those of the overexpression lines L6 and L8, indicating that the photosynthesis of alfalfa SY4D was greatly affected by salt stress.

3.12 Analysis of the biochemical indicators of salt tolerance and the contents of K⁺ and Na⁺ in root and leaf tissues of alfalfa plants overexpressing the target gene

The results are shown in Figure 14. After 15d of salt treatment, the proline and malondialdehyde contents of the overexpressing lines increased significantly, but the malondialdehyde content of the

control was significantly higher than that of the overexpressing lines, indicating that salt stress caused greater damage to the cell membrane of the wild type alfalfa SY4D cells. The superoxide dismutase and catalase activities decreased significantly, while that of peroxidase increased significantly in the overexpressing lines compared to the control, indicating that the antioxidant capacity of the transgenic lines was significantly improved. Abscisic acid content in the roots of both control and transgenic plants was significantly higher after 24 h of salt treatment than under normal conditions, but the abscisic acid in the roots and leaves of transgenic plants was significantly higher than that in the control. Prior to salt stress, the abscisic acid (ABA) content in the transgenic plants was significantly higher than that of the control.





As shown in Figure 15, the Na^+ content in the roots and leaves of transgenic plants was significantly lower, but their K^+/Na^+ ratio was significantly higher under salt stress compared to control. The K^+ content in the roots of transgenic L6 and L8 plants and the K^+

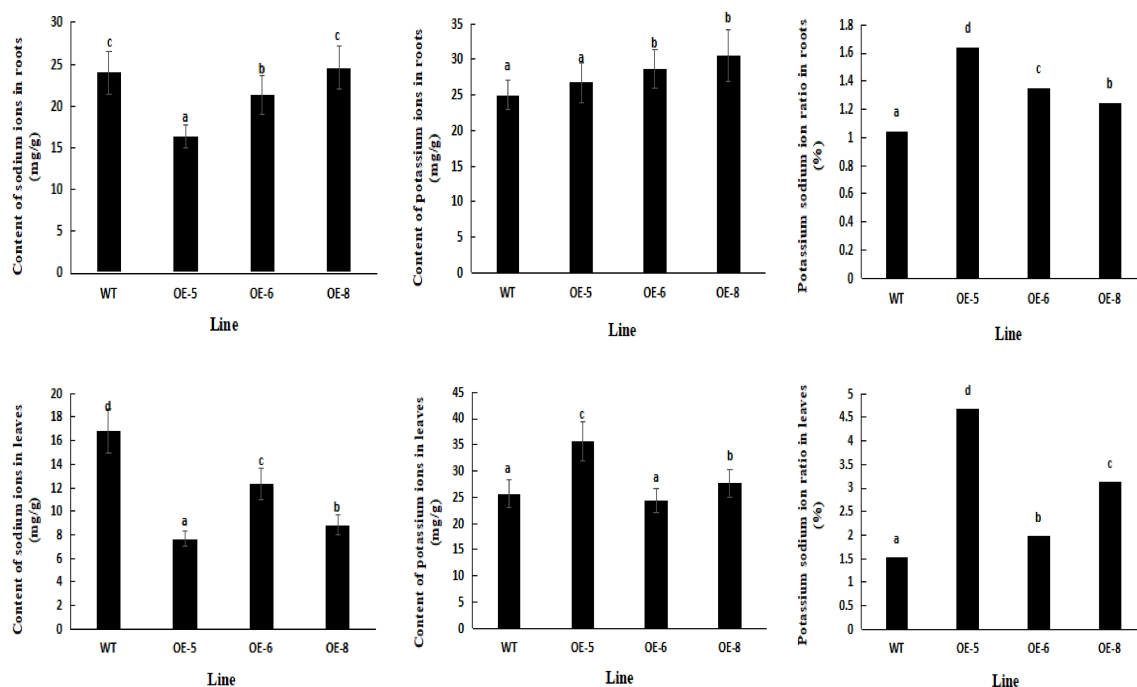


FIGURE 8

The expression levels of 15 candidate genes in root and leaf under different stress. The blue group is the blank control group, and the red group is 50 In the mmol/L NaCl salt treatment group, the gray was 100 mmol/L In the NaCl salt treatment group, yellow was 150 mmol/L NaCl salt treatment group. In bar charts, "abc" denotes the results of significance analysis.

content in the leaves of L5 and L8 plants under salt stress was significantly higher than that of the control. This indicated that the transgenic plants could effectively reduce the uptake of Na⁺ by the roots and leaves, maintain the stability of K⁺ in the root and leaf tissues, maintain the internal homeostasis of K⁺/Na⁺, and reduce the toxicity of Na⁺.

4 Discussion

This study conducted a comprehensive and systematic bioinformatics analysis of the alfalfa *NAC* gene members. Considering the possibility that salt-tolerant genes are also responsive to other abiotic stresses, we adopted a screening strategy with salt stress as the main stress and saline and alkaline co-stress as the secondary stress based on the transcriptomic data of alkali stress and saline and alkaline co-stress. According to the analysis of promoter cis-acting elements and salt tolerance of the subfamilies of the alfalfa *NAC* gene family, subfamily V was found to be mostly associated with stress tolerance. Moreover, based on the expression levels of the candidate genes, the *MsNAC40* gene, a member of subfamily V, was selected for subsequent analysis.

We screened 114 alfalfa *NAC* genes based on the Zhongmu No.1 genome using the Hidden Markov Model and conducted phylogenetic analyses of the alfalfa *NAC* gene family members. He et al. (2022) screened 421 alfalfa *NAC* genes from the Xinjiang Daye genome and identified 25, 42 and 47 alfalfa genes responsive to cold, drought and salt stress, respectively, via transcriptomic and qPCR analyses (He et al.,

2022). In this study, phylogenetic analysis of the alfalfa *NAC* gene family confirmed that alfalfa is a homotetraploid. The *NAC* genes in the Zhongmu No.1 genome were distributed across eight chromosomes, while in the Xinjiang Large-Leaf genome, they were distributed across 32 chromosomes, reflecting the fact that the Xinjiang Large-Leaf alfalfa genome consists of four haploid genome sets, whereas Zhongmu No.1 consists of only one. Given this fundamental genomic difference, direct comparisons of *NAC* gene family distribution between the two genomes may have limited significance. To further advance research on the *NAC* gene family in alfalfa, the [Supplementary Materials](#) of this paper provide a comparative analysis of *NAC* gene members between Zhongmu No.1 and Xinjiang Large-Leaf alfalfa. The observed gene distribution aligns with patterns seen in other polyploid species, such as cotton (Wu et al., 2013), and may reflect evolutionary pressures and functional adaptations of polyploid plants under abiotic stress conditions. Finally, based on the expression levels of candidate genes, *MsNAC40* from subfamily V was selected as the focus for further research.

The primary structure, tertiary structure, hydrophilicity, and prediction analysis of phosphorylation sites of the protein *MsNAC40* showed that *MsNAC40* has 329 amino acids and a molecular weight of 3.70 KDa, with the NAM a conserved structural domain, and is a hydrophilic protein with no transmembrane structure. Tissue-specific expression of *MsNAC40* was analysed, and it was found that the relative expression level of *MsNAC40* was the highest in the roots and the lowest in flowers and was expressed to different degrees in the stems, leaves and branches of alfalfa plants.

The fresh weight and plant height of L8 were significantly higher than that of the control, while the plant heights of L5 and

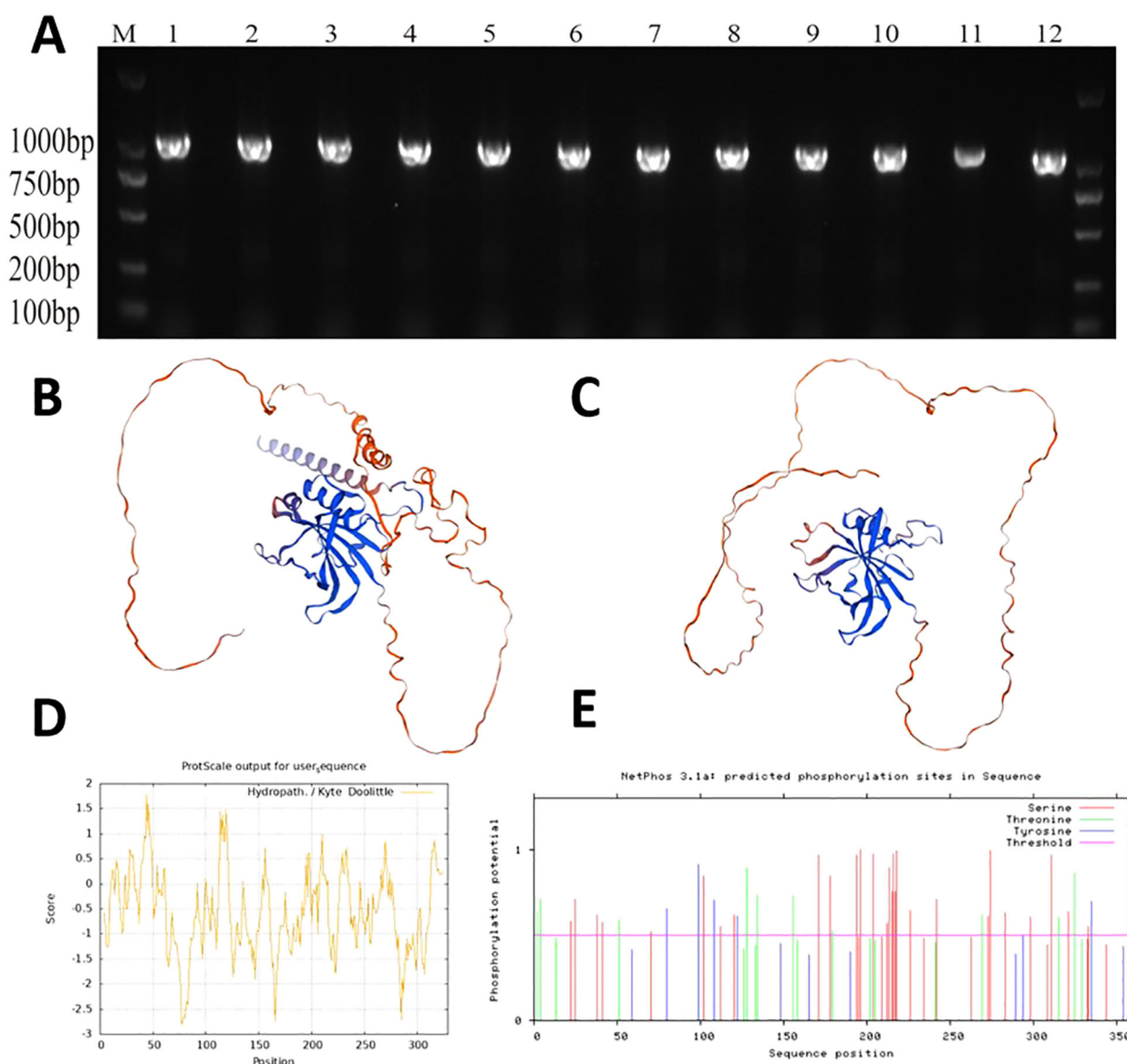


FIGURE 9

Analysis of *MsNAC40* Gene and Protein Structure and Function Prediction. (A) is the clone of *MsNAC40*, (B) is the predicted tertiary structure of *OsNAC1* protein, (C) is the predicted tertiary structure of *MsNAC40* protein, (D) is the predicted hydrophilicity of *MsNAC40*, and (E) is the predicted phosphorylation site of *MsNAC40*.

L6 were significantly higher than those of the control. This indicated that the salt tolerance of *MsNAC40*-overexpressing plants was improved compared with that of the control. The above-ground biomass of L5 and L6 did not differ significantly from the control, probably because the duration of salt treatment was too short and *MsNAC40* didn't play the antioxidation role in the leaves directly, resulting in the wilting and yellowing of the leaves of the overexpression plants, similar to the control leaves.

Furthermore, the photosynthetic indexes of the plant, such as net photosynthetic rate, stomatal conductance, and transpiration rate, were significantly decreased after the salt treatment compared with that before the treatment, indicating that salt stress affected the photosynthesis of the plants. The photosynthetic indexes of *MsNAC40*-overexpressing plants, except for the transpiration rate

of the L5 line, were significantly higher than that of the control. It is likely that variations in expression levels among different plants contribute to the differences in net photosynthetic rate responses observed in L5 compared to L6 and L8, suggesting that the photosynthesis of *MsNAC40*-overexpressing plants was less affected by the salt stress. The conductivity of the leaves of the control group was significantly higher than that of the *MsNAC40*-overexpressing plants, indicating that the leaves of the control group were more damaged by salt stress and that the *MsNAC40*-overexpressing plants were more salt-tolerant than the control plants.

Malondialdehyde and proline contents were significantly elevated in the plant leaves after salt stress; however, the accumulation of malondialdehyde was significantly higher in the

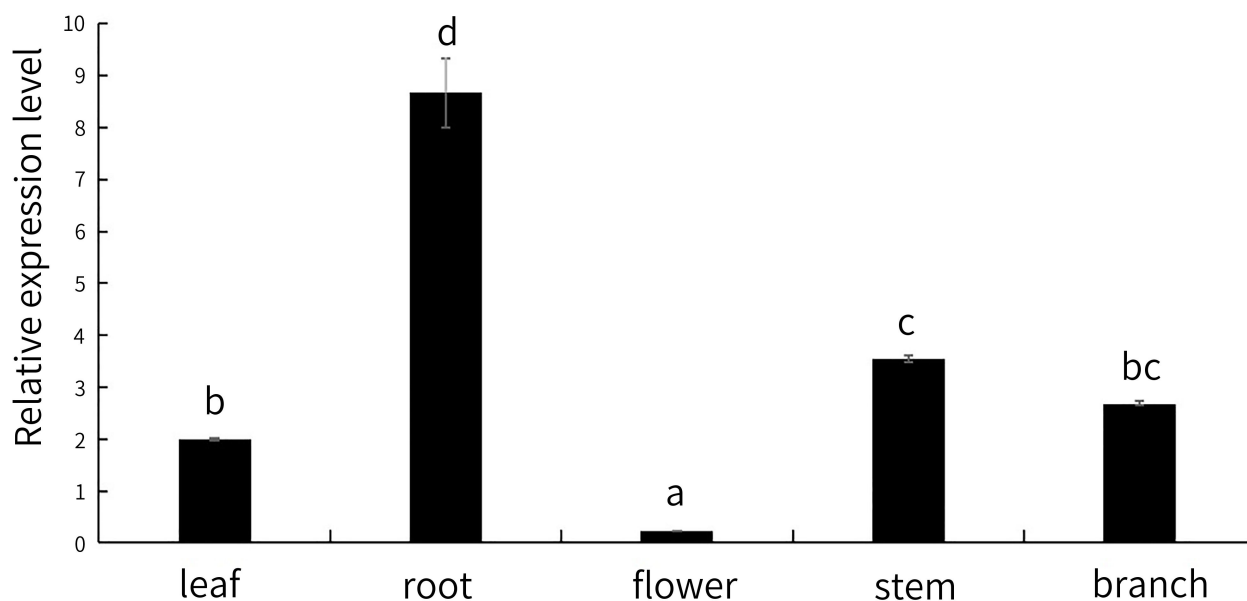


FIGURE 10

Expression levels of *MsNAC40* in different tissues. In bar charts, "abc" denotes the results of significance analysis.

control leaves than in those of the *MsNAC40*-overexpressing plants. The proline content of the control plants did not differ significantly from that of *MsNAC40*-overexpressing L8 lines and was significantly higher in L5 and L6 plants. The variation in response to proline content between L8 and L5/L6 might be attributed to differences in expression levels across individual plants. Malondialdehyde is one of the important products of membrane lipid peroxidation, and its production can also exacerbate membrane damage (Jones, 2007), indicating that the leaves of *MsNAC40*-overexpressing plants were less damaged by salt stress. The physiological significance of proline accumulation in plants under salt stress is conflicted. One view is that proline accumulation

can increase plant tolerance to osmotic stress because it can regulate the ionic balance in plants, thus maintaining the balance of intra- and extracellular concentrations and reducing cellular water loss (Wang et al., 2015). Nonetheless, proline can also protect biomolecules such as proteins and membrane lipids and enhance plant adaptation to other stresses (Ben Rejeb et al., 2012; Ghosh et al., 2022). However, judging from the significantly increased proline content in the control and overexpression plants after salt stress in this study, both control and overexpression plants were affected by salt stress, with slight differences them, suggesting that *MsNAC40* may not be associated with free proline accumulation in alfalfa.

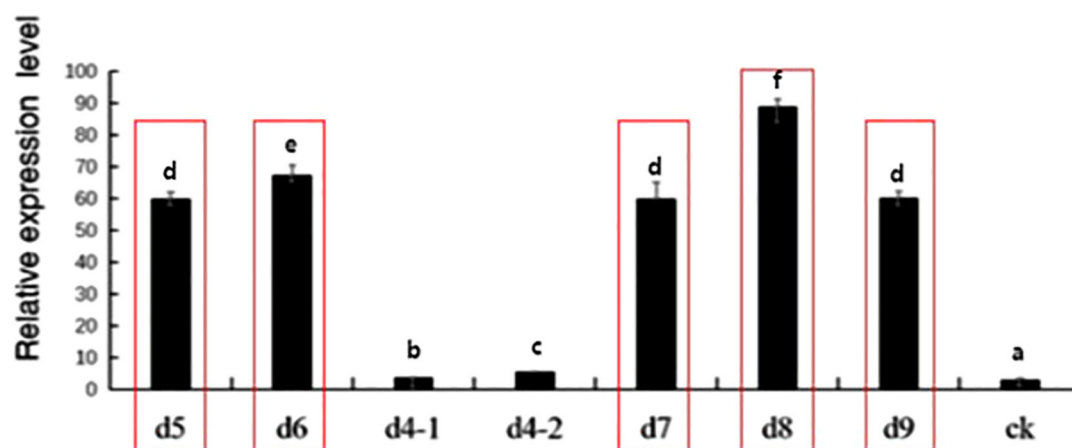


FIGURE 11

Expression level of *MsNAC40* in overexpressed positive plants. M is 1500bp DNA Marker, WT is pCAMIBA3301 vector with empty plasmid as the template, and 1 to 35 are the 35 *MsNAC40*-overexpressing seedlings positive for the target gene. In bar charts, "abc" denotes the results of significance analysis.

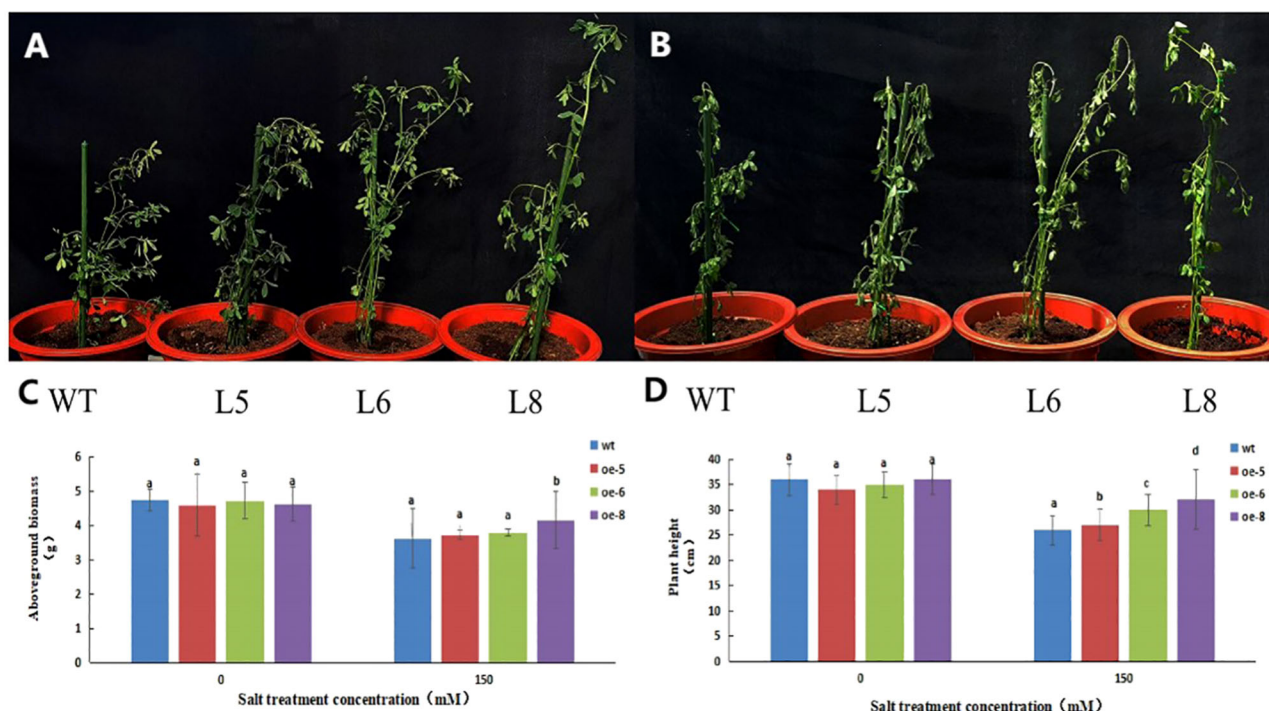


FIGURE 12

Phenotypic analysis of *MsNAC40* overexpressed plants under salt stress. (A) is the growth state under normal condition; (B) is the growth state after salt treatment for 15d; (C) is the fresh weight of the plant; (D) is the plant height. In bar charts, "abc" denotes the results of significance analysis.

MsNAC40 was found to contain more cis-acting elements related to abscisic acid synthesis, and the abscisic acid content was increased in the roots and leaves of *MsNAC40*-overexpressing plants compared to the control, but the content was decreased in the leaves after 24 h of salt stress. Studies have shown that all abiotic stresses can induce a rapid increase in abscisic acid content in plants, thus increasing their stress tolerance (Dong et al., 2019). Abscisic acid induces the resynthesis of

plant enzymes, increases plant salt resistance (Kou et al., 2021), significantly reduces the organellar ultrastructure damage caused by high temperatures and other adversities, and increases the stability of organelles (Lv et al., 2022; Tao et al., 2022). Thus, identifying the key genes associated with the synthesis of resistance hormones such as abscisic acid could help clarify how *MsNAC40* increases abscisic acid content in plants.

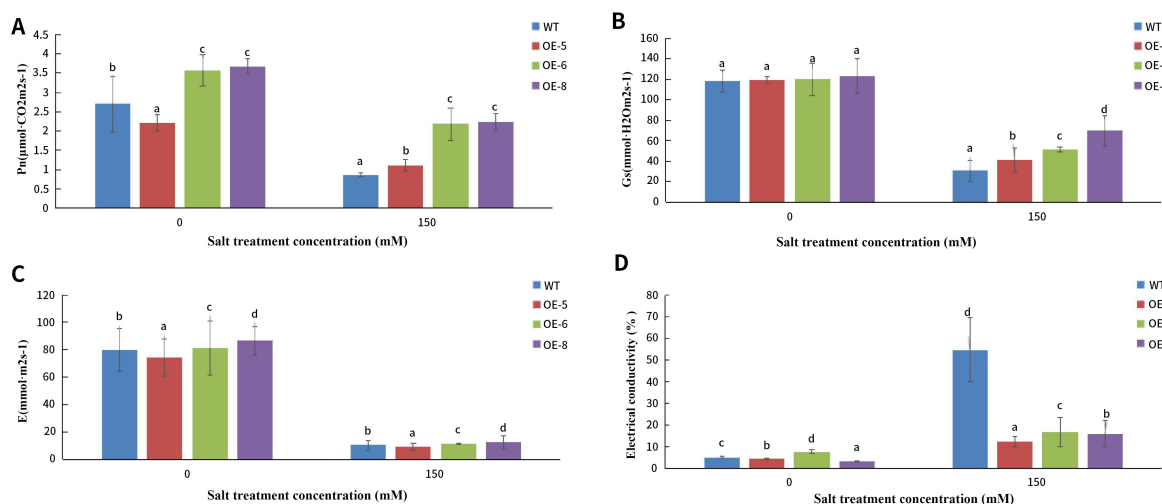


FIGURE 13

Physiological indices of *MsNAC40* overexpressed plants under salt stress. (A) is the measurement of net photosynthetic rate of leaves; (B) is the measurement of stomatal conductance of leaves; (C) is the measurement of transpiration rate of leaves; (D) is the measurement of electrical conductivity of leaves. In bar charts, "abc" denotes the results of significance analysis.

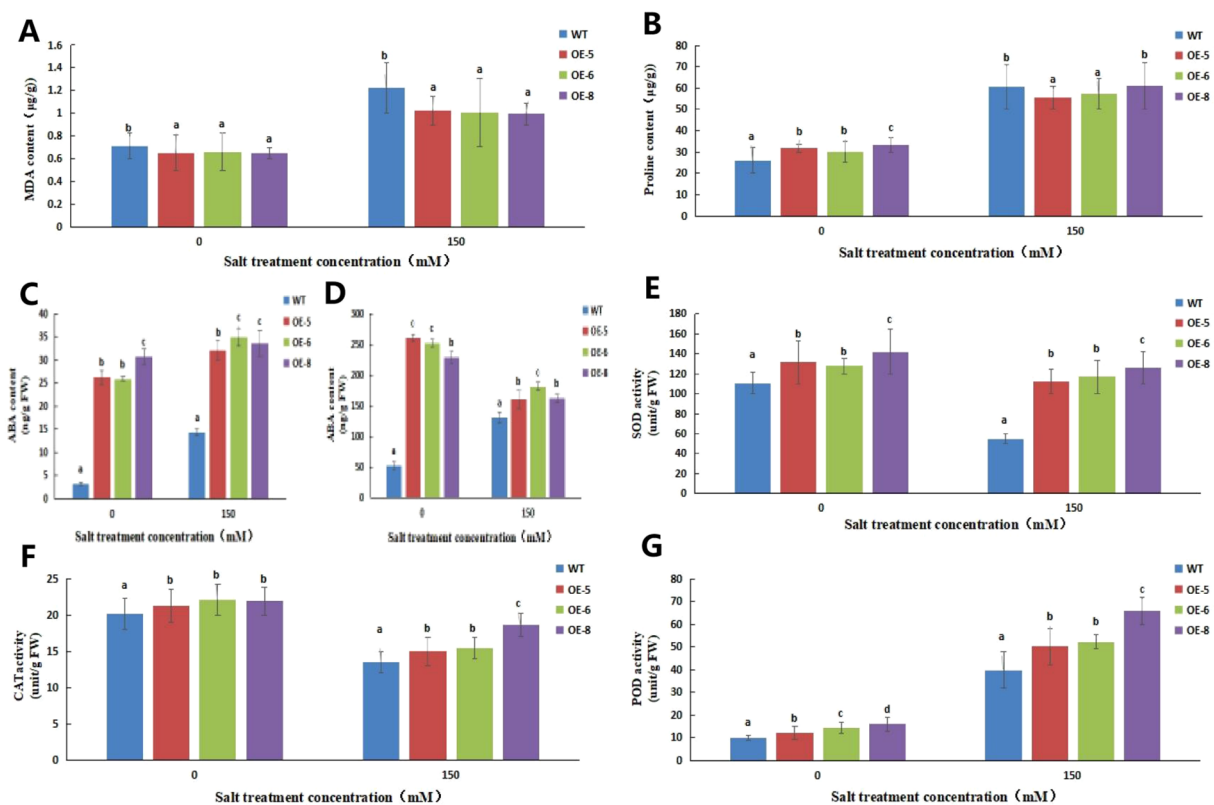


FIGURE 14

Analysis of biochemical indices of *MsNAC40* overexpressed plants under salt stress. (A) is the malondialdehyde content in the leaves, (B) is the proline content in the leaves, (C) is the abscisic acid content in the roots, (D) is the abscisic acid content in the leaves, (E) is the superoxide dismutase activity, (F) is the catalase activity, (G) is the peroxidase activity. In bar charts, "abc" denotes the results of significance analysis.

We analysed the activities of three antioxidant enzymes, superoxide dismutase, peroxide, and hydrogen peroxide, and found a significant decrease in superoxide dismutase activity after 15 d of salt treatment compared to the control treatment. This might have been because the salt treatment duration was too short, resulting in the production of other osmotic substances in the plants, thus inhibiting the large accumulation of intracellular reactive oxygen species (Anjum et al., 2015; Gill et al., 2015). The prolonged duration of salt stress increased the accumulation of reactive oxygen species, further increasing superoxide dismutase activity in the leaves (Gharsallah et al., 2016). Jin suggests that salt stress significantly increases peroxidase activity in salt-tolerant plants, thereby enhancing both salt tolerance and antioxidant responses in soybeans (Jin et al., 2019). Peroxidase activity of the *MsNAC40*-overexpressing lines was significantly higher than that of the control, suggesting that the *MsNAC40*-overexpressing lines had better salt tolerance than the control. As one of the major scavengers of cellular reactive oxygen species, catalase plays an important role in plant salt tolerance (Wang et al., 2023; Zhang et al., 2013). Rout and Shaw (2001) concluded that enhanced catalase activity is closely related to plant salt tolerance (Rout and Shaw, 2001). The results of the present study revealed a significant

decrease in catalase activity in the leaves of control and overexpression plants after 15 d of salt treatment compared to the normal conditions, suggesting a lower association between *MsNAC40* and catalase synthesis. However, since determining the effect of antioxidant enzyme activity only in the leaves is not comprehensive enough, enzyme activity should be further analysed in different parts of the plants under salt stress.

A balanced ratio of mineral nutrients to sodium in plants under salt-stressed environments is a physiological manifestation of plant salt tolerance, and a higher K^+/Na^+ ratio is one of the important indicators of plant salt tolerance (Bassil et al., 2011; Blumwald, 2000; Deinlein et al., 2014; Muchate et al., 2016; van Zelm et al., 2020). This study showed that the Na^+ content in the roots and leaves of the *MsNAC40*-overexpressing plants was significantly lower than that of the control plants under salt stress, and the K^+/Na^+ was significantly higher. This indicated that *MsNAC40* could prevent Na^+ from entering the cells, alleviate the competitive effects between K^+ and Na^+ , promote the uptake of K^+ , reduce the ionic toxicity of Na^+ , and maintain the internal homeostasis ratio of K^+/Na^+ in alfalfa. Overall, the salt tolerance capacity of the *MsNAC40*-overexpressing plants was enhanced under salt stress.

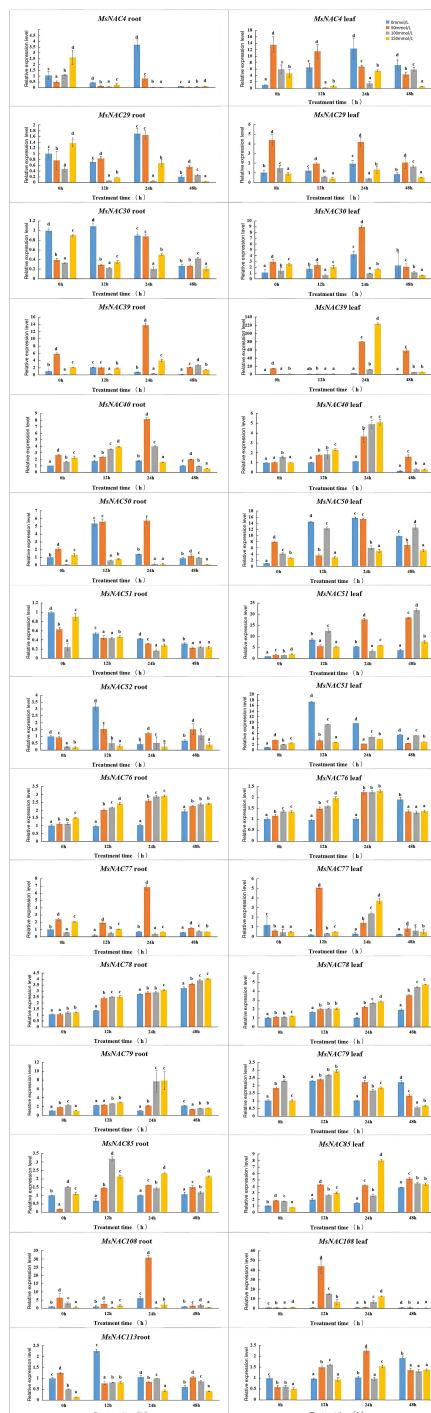


FIGURE 15

The contents of K^+ and Na^+ in *MsNAC40* overexpressed plants under salt stress. (A) is the content of Na^+ in the root, (B) is the content of K^+ in the root, (C) is the content of K^+/Na^+ in the root, (D) is the content of Na^+ in the leaf, (E) is the content of K^+ in the leaf, (F) is the content of K^+/Na^+ in the leaf.

5 Conclusions

This study identified 114 NAC gene family members from the Zhongmu No.1 genome for the first time and classified the genes

into 13 subclasses (I to XIII). All identified genes are hydrophilic proteins. Most genes contain the conserved NAM domain, and family members generally exhibit 8 conserved motifs. All genes contain introns, with subfamilies V, VIII, and XII showing simpler structures and higher intragenic similarity. The distribution of genes across chromosomes is relatively even. The promoter regions are rich in cis-elements related to light response, hormone regulation, and abiotic stress, suggesting that the V subfamily may play a role in regulating ABA and anaerobic stress responses.

The open reading frame of the *MsNAC40* gene was 990 bp, encoding a 329 amino acids-long hydrophilic protein without a transmembrane structure, with a molecular weight of 3.70KDa and a NAM conserved structural domain. The physiological indexes of the *MsNAC40*-overexpressing plants, except conductivity, and their biochemical indexes, except malondialdehyde content, were significantly higher than in the control. Similarly, the K^+/Na^+ ratio in the roots and leaves of the *MsNAC40*-overexpressing plants was significantly higher than in the control under salt stress. In conclusion, the salt tolerance of *MsNAC40*-overexpressing plants was improved under salt stress compared with that of the wild-type alfalfa SY4D.

Data availability statement

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

Author contributions

ZL: Conceptualization, Writing – original draft, Writing – review & editing, Data curation, Formal Analysis, Investigation, Methodology. QY: Formal Analysis, Writing – original draft, Conceptualization, Data curation, Investigation, Methodology, Writing – review & editing. YM: Data curation, Investigation, Methodology, Writing – original draft. FM: Data curation, Investigation, Project administration, Writing – review & editing. LM: Conceptualization, Data curation, Project administration, Supervision, Writing – review & editing. HZ: Methodology, Validation, Writing – review & editing. SL: Data curation, Investigation, Project administration, Supervision, Writing – original draft. KS: Data curation, Investigation, Project administration, Validation, Writing – original draft. ZW: Conceptualization, Project administration, Writing – review & editing. GY: Data curation, Validation, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This study was supported by the National Nature Science Foundation of China (U1906201), Shandong Forage Research System (SDAIT-23-01), China

Agriculture Research System (CARS-34), the First Class Grassland Science Discipline Program of Shandong Province (1619002), China and the Foundation Project of Shandong Natural Science Foundation (ZR2022MC031), and the Shandong Province Key Research and Development Plan (2021SFGC0303, 2023LZGCQY022).

Acknowledgments

The authors would like to thank Professors Guofeng Yang, Zengyu Wang, and Juan Sun (Professor of Grassland Science, Qingdao Agricultural University) for their help in analyzing the data and writing the manuscript. We are also grateful for the research funding provided by the College of Grassland Science of Qingdao Agricultural University and the experimental help provided by Berry Hekang (Beijing, China). We would like to thank MogoEdit (<https://www.mogoedit.com>) for its English editing during the preparation of this manuscript.

Conflict of interest

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

References

- Anjum, N. A., Sofo, A., Scopa, A., Roychoudhury, A., Gill, S. S., Iqbal, M., et al. (2015). Lipids and proteins-major targets of oxidative modifications in abiotic stressed plants. *Environ. Sci. Pollut. Res.* 22, 4099–4121. doi: 10.1007/s11356-014-3917-1
- Balazadeh, S., Siddiqui, H., Allu, A. D., Matallana-Ramirez, L. P., Caldana, C., Mehrnia, M., et al. (2010). A gene regulatory network controlled by the NAC transcription factor ANAC092/AtNAC2/ORE1 during salt-promoted senescence. *Plant J.* 62, 250–264. doi: 10.1111/j.1365-3113.2010.04151.x
- Bassil, E., Ohto, M.-A., Esumi, T., Tajima, H., Zhu, Z., Cagnac, O., et al. (2011). The Arabidopsis intracellular Na⁺/H⁺ antiporters NHX5 and NHX6 are endosome associated and necessary for plant growth and development. *Plant Cell* 23, 224–239. doi: 10.1105/tpc.110.079426
- Ben Rejeb, K., Abdely, C., and Savouré, A. (2012). Proline, a multifunctional amino-acid involved in plant adaptation to environmental constraints. *Biol. Aujourd'hui* 206, 291–299. doi: 10.1051/jbio/2012030
- Blumwald, E. (2000). Sodium transport and salt tolerance in plants. *Curr. Opin. Cell Biol.* 12, 431–434. doi: 10.1016/S0955-0674(00)00112-5
- Chen, C. J., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y. H., et al. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. *Mol. Plant* 13, 1194–1202. doi: 10.1016/j.molp.2020.06.009
- Christianson, J. A., Wilson, I. W., Llewellyn, D. J., and Dennis, E. S. (2009). The low-oxygen-induced NAC domain transcription factor ANAC102 affects viability of Arabidopsis seeds following low-oxygen treatment. *Plant Physiol.* 149, 1724–1738. doi: 10.1104/pp.108.131912
- Deinlein, U., Stephan, A. B., Horie, T., Luo, W., Xu, G. H., and Schroeder, J. I. (2014). Plant salt-tolerance mechanisms. *Trends Plant Sci.* 19, 371–379. doi: 10.1016/j.tplants.2014.02.001
- Diao, P. F., Chen, C., Zhang, Y. Z., Meng, Q. W., Lv, W., and Ma, N. N. (2020). The role of NAC transcription factor in plant cold response. *Plant Signaling Behavior*. 15 (9), 1785668. doi: 10.1080/15592324.2020.1785668
- Dong, W., Liu, X., Lv, J., Gao, T., and Song, Y. (2019). The expression of alfalfa MsPP2CA1 gene confers ABA sensitivity and abiotic stress tolerance on Arabidopsis thaliana. *Plant Physiol. Biochem.* 143, 176–182. doi: 10.1016/j.plaphy.2019.09.004
- Duval, M., Hsieh, T. F., Kim, S. Y., and Thomas, T. L. (2002). Molecular characterization of AtNAM: a member of the Arabidopsis NAC domain superfamily. *Plant Mol. Biol.* 50, 237–248. doi: 10.1023/A:1016028530943
- Fang, L. F., Liu, T., Li, M. Y., Dong, X. M., Han, Y. L., Xu, C. Z., et al. (2024). MODMS: a multi-omics database for facilitating biological studies on alfalfa (*Medicago sativa* L.). *Horticulture Res.* 11 (1), uhad245. doi: 10.1093/hr/uhad245
- Gharsallah, C., Fakhfakh, H., Grubb, D., and Gorsane, F. (2016). Effect of salt stress on ion concentration, proline content, antioxidant enzyme activities and gene expression in tomato cultivars. *AoB Plants* 8, plw055. doi: 10.1093/aobpla/plw055
- Ghosh, U. K., Islam, M. N., Siddiqui, M. N., Cao, X., and Khan, M. A. R. (2022). Proline, a multifaceted signalling molecule in plant responses to abiotic stress: understanding the physiological mechanisms. *Plant Biol. (Stuttg.)* 24, 227–239. doi: 10.1111/plb.13363
- Gill, S. S., Anjum, N. A., Gill, R., Yadav, S., Hasanuzzaman, M., Fujita, M., et al. (2015). Superoxide dismutase-mentor of abiotic stress tolerance in crop plants. *Environ. Sci. Pollut. Res.* 22, 10375–10394. doi: 10.1007/s11356-015-4532-5
- Guo, H. S., Xie, Q., Fei, J. F., and Chua, N. H. (2005). MicroRNA directs mRNA cleavage of the transcription factor NAC1 to downregulate auxin signals for Arabidopsis lateral root development. *Plant Cell* 17, 1376–1386. doi: 10.1105/tpc.105.030841
- Hao, Y. J., Wei, W., Song, Q. X., Chen, H. W., Zhang, Y. Q., Wang, F., et al. (2011). Soybean NAC transcription factors promote abiotic stress tolerance and lateral root formation in transgenic plants. *Plant J.* 68, 302–313. doi: 10.1111/j.1365-3113.2011.04687.x
- He, F., Long, R., Wei, C., Zhang, Y., Li, M., Kang, J., et al. (2022). Genome-wide identification, phylogeny and expression analysis of the SPL gene family and its important role in salt stress in *Medicago sativa* L. *BMC Plant Biol.* 22, 295. doi: 10.1186/s12870-022-03678-7
- Hu, R., Qi, G., Kong, Y., Kong, D., Gao, Q., and Zhou, G. (2010). Comprehensive analysis of NAC domain transcription factor gene family in *Populus trichocarpa*. *BMC Plant Biol.* 10, 145. doi: 10.1186/1471-2229-10-145
- Hu, H., You, J., Fang, Y., Zhu, X., Qi, Z., and Xiong, L. (2008). Characterization of transcription factor gene SNAC2 conferring cold and salt tolerance in rice. *Plant Mol. Biol.* 67, 169–181. doi: 10.1007/s11103-008-9309-5
- Jensen, M. K., Kjaersgaard, T., Petersen, K., and Skriver, K. (2010). NAC genes: time-specific regulators of hormonal signaling in Arabidopsis. *Plant Signal Behav.* 5, 907–910. doi: 10.4161/psb.5.7.12099
- Jeong, J. S., Kim, Y. S., Baek, K. H., Jung, H., Ha, S. H., Do Choi, Y., et al. (2010). Root-specific expression of OsNAC10 improves drought tolerance and grain yield in

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.2025.1461735/full#supplementary-material>

SUPPLEMENTARY FIGURE 1

Map of overexpressed vector pCAMIBA3301.MsNAC40.

SUPPLEMENTARY FIGURE 2

Diagram of genetic transformation of Alfalfa. (A) shows the state of co-cultured leaves in the dark, (B) and (C) show the callus state in selected medium SH3a, (D) and (E) show the state of green buds emerging from callus on MSBK medium, (E) and (F) show the state of rooting of callus on SH9a medium. In bar charts, “abc” denotes the results of significance analysis.

SUPPLEMENTARY FIGURE 3

Identification of MsNAC40 in overexpressed positive plants. M is 1500bp DNA Marker, WT is pCAMIBA3301 vector with empty plasmid as the template, and 1 to 35 are the 35 MsNAC40-overexpressing seedlings positive for the target gene.

- rice under field drought conditions. *Plant Physiol.* 153, 185–197. doi: 10.1104/pp.110.154773
- Jin, T., Sun, Y. Y., Zhao, R. R., Shan, Z., Gai, J. Y., and Li, Y. (2019). Overexpression of peroxidase gene *gsPRX9* confers salt tolerance in soybean. *Int. J. Mol. Sci.* 20 (15), 3745. doi: 10.3390/ijms20153745
- Jones, M. R. (2007). Lipids in photosynthetic reaction centres: structural roles and functional holes. *Prog. Lipid Res.* 46, 56–87. doi: 10.1016/j.plipres.2006.06.001
- Kim, S. G., Kim, S. Y., and Park, C. M. (2007). A membrane-associated NAC transcription factor regulates salt-responsive flowering via *FLOWERING LOCUS T* in *Arabidopsis*. *Planta*. 226, 647–654. doi: 10.1007/s00425-007-0513-3
- Kim, Y. S., Kim, S. G., Park, J. E., Park, H. Y., Lim, M. H., Chua, N. H., et al. (2006). A membrane-bound NAC transcription factor regulates cell division in *Arabidopsis*. *Plant Cell*. 18, 3132–3144. doi: 10.1105/tpc.106.043018
- Ko, J. H., Yang, S. H., Park, A. H., Lerouxel, O., and Han, K. H. (2007). ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*. *Plant J.* 50, 1035–1048. doi: 10.1111/j.1365-3113.2007.03109.x
- Kou, X., Yang, S., Chai, L., Wu, C., Zhou, J., Liu, Y., et al. (2021). Absciscic acid and fruit ripening: Multifaceted analysis of the effect of absciscic acid on fleshy fruit ripening. *Scientia Horticulturae*. 281, 109999. doi: 10.1016/j.scienta.2021.109999
- Le, D. T., Nishiyama, R., Watanabe, Y., Mochida, K., Yamaguchi-Shinozaki, K., Shinozaki, K., et al. (2011). Genome-wide survey and expression analysis of the plant-specific NAC transcription factor family in soybean during development and dehydration stress. *DNA Res.* 18, 263–276. doi: 10.1093/dnares/dsr015
- 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, S., Wang, N., Ji, D., Zhang, W., Wang, Y., Yu, Y., et al. (2019). A *gmSIN1/gmNCED3s/gmRbohBs* feed-forward loop acts as a signal amplifier that regulates root growth in soybean exposed to salt stress. *Plant Cell*. 31, 2107–2130. doi: 10.1105/tpc.18.00662
- Ling, L., Song, L., Wang, Y., and Guo, C. (2017a). Genome-wide analysis and expression patterns of the NAC transcription factor family in *Medicago truncatula*. *Physiol. Mol. Biol. Plants*. 23, 343–356. doi: 10.1007/s12298-017-0421-3
- Lu, M., Ying, S., Zhang, D. F., Shi, Y. S., Song, Y. C., Wang, T. Y., et al. (2012). A maize stress-responsive NAC transcription factor, *ZmSNAC1*, confers enhanced tolerance to dehydration in transgenic *Arabidopsis*. *Plant Cell Rep.* 31, 1701–1711. doi: 10.1007/s00299-012-1284-2
- Lv, J. H., Dong, T. Y., Zhang, Y. P., Ku, Y., Zheng, T., Jia, H. F., et al. (2022). Metabolomic profiling of brassinolide and absciscic acid in response to high-temperature stress. *Plant Cell Rep.* 41, 935–946. doi: 10.1007/s00299-022-02829-2
- Muchate, N. S., Nikalje, G. C., Rajurkar, N. S., Suprasanna, P., and Nikam, T. D. (2016). Plant salt stress: adaptive responses, tolerance mechanism and bioengineering for salt tolerance. *Botanical Review*. 82, 371–406. doi: 10.1007/s12229-016-9173-y
- Nuruzzaman, M., Manimekalai, R., Sharoni, A. M., Satoh, K., Kondoh, H., Ooka, H., et al. (2010). Genome-wide analysis of NAC transcription factor family in rice. *Gene*. 465, 30–44. doi: 10.1016/j.gene.2010.06.008
- Ooka, H., Satoh, K., Doi, K., Nagata, T., Otomo, Y., Murakami, K., et al. (2003). Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. *DNA Res.* 10, 239–247. doi: 10.1093/dnares/10.6.239
- Rout, N. P., and Shaw, B. P. (2001). Salt tolerance in aquatic macrophytes: possible involvement of the antioxidative enzymes. *Plant Sci.* 160, 415–423. doi: 10.1016/S0168-9452(00)00406-4
- Rushton, P. J., Bokowiec, M. T., Han, S., Zhang, H., Brannock, J. F., Chen, X., et al. (2008). Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant Physiol.* 147, 280–295. doi: 10.1104/pp.107.114041
- Shahnejat-Bushehri, S., Allu, A. D., Mehterov, N., Thirumalaikumar, V. P., Alseekh, S., Fernie, A. R., et al. (2017). *Arabidopsis* NAC transcription factor *JUNGBRUNNEN1* exerts conserved control over gibberellin and brassinosteroid metabolism and signaling genes in tomato. *Front. Plant Sci.* 8, 214. doi: 10.3389/fpls.2017.00214
- Shao, H. B., Chu, L. Y., Lu, H. Y., Qi, W. C., Chen, X., Liu, J., et al. (2019). Towards sustainable agriculture for the salt-affected soil. *Land Degradation Dev.* 30, 574–579. doi: 10.1002/ldr.v30.5
- Sperotto, R. A., Ricachenevsky, F. K., Duarte, G. L., Boff, T., Lopes, K. L., Sperb, E. R., et al. (2009). Identification of up-regulated genes in flag leaves during rice grain filling and characterization of *OsNAC5*, a new ABA-dependent transcription factor. *Planta*. 230, 985–1002. doi: 10.1007/s00425-009-1000-9
- Tao, Z. Q., Yan, P., Zhang, X. P., Wang, D. M., Wang, Y. J., Ma, X. L., et al. (2022). Physiological mechanism of absciscic acid-induced heat-tolerance responses to cultivation techniques in wheat and maize-review. *Agronomy-Basel*. 12 (7), 1579. doi: 10.3390/agronomy12071579
- Tran, L. S., Nakashima, K., Sakuma, Y., Simpson, S. D., Fujita, Y., Maruyama, K., et al. (2004). Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive cis-element in the early responsive to dehydration stress 1 promoter. *Plant Cell*. 16, 2481–2498. doi: 10.1105/tpc.104.022699
- van Zelm, E., Zhang, Y., and Testerink, C. (2020). Salt tolerance mechanisms of plants. *Annu. Rev. Plant Biol.* 71, 403–433. doi: 10.1146/annurev-arplant-050718-100005
- Wan, W. F., Liu, Q., Zhang, C. H., Li, K., Sun, Z., Li, Y. J., et al. (2023). Alfalfa growth and nitrogen fixation constraints in salt-affected soils are in part offset by increased nitrogen supply. *Front. Plant Science*. 14. doi: 10.3389/fpls.2023.1126017
- Wang, H. Y., Tang, X. L., Wang, H. L., and Shao, H. B. (2015). Proline accumulation and metabolism-related genes expression profiles in *Kosteletzkya virginica* seedlings under salt stress. *Front. Plant Science*. 6. doi: 10.3389/fpls.2015.00792
- Wang, Y., Yi, Y. T., Liu, C., Zheng, H. P., Huang, J., Tian, Y., et al. (2023). Dephosphorylation of CatC at Ser-18 improves salt and oxidative tolerance via promoting its tetramerization in rice. *Plant Science*. 329, 111597. doi: 10.1016/j.plantsci.2023.111597
- Wang, L., Zhang, W., Wang, L., Zhang, X. C., Li, X., and Rao, Z. (2010). Crystal structures of NAC domains of human nascent polypeptide-associated complex (NAC) and its α NAC subunit. *Protein Cell*. 1, 406–416. doi: 10.1007/s13238-010-0049-3
- Wu, Y. X., Chen, J. H., He, Q. L., and Zhu, S. J. (2013). Parental origin and genomic evolution of tetraploid *Gossypium* species by molecular marker and GISH analyses. *Caryologia* 66, 368–374. doi: 10.1080/00087114.2013.857830
- Yuan, X., Wang, H., Cai, J. T., Li, D. Y., and Song, F. M. (2019). NAC transcription factors in plant immunity. *Phytopathol. Res.* 1, 3. doi: 10.1186/s42483-018-0008-0
- Zhang, B. G., Liu, K. D., Zheng, Y., Wang, Y. X., Wang, J. X., and Liao, H. (2013). Disruption of *atWNK8* enhances tolerance of *Arabidopsis* to salt and osmotic stresses via modulating proline content and activities of catalase and peroxidase. *Int. J. Mol. Sci.* 14, 7032–7047. doi: 10.3390/ijms14047032
- Zhang, X. X., Sun, Y., Qiu, X., Lu, H., Hwang, I., and Wang, T. Z. (2022). Tolerant mechanism of model legume plant *Medicago truncatula* to drought, salt, and cold stresses. *Front. Plant Science*. 13. doi: 10.3389/fpls.2022.847166
- Zheng, X., Chen, B., Lu, G., and Han, B. (2009). Overexpression of a NAC transcription factor enhances rice drought and salt tolerance. *Biochem. Biophys. Res. Commun.* 379, 985–989. doi: 10.1016/j.bbrc.2008.12.163
- Zhu, Q. K., Zou, J. X., Zhu, M. L., Liu, Z. B., Feng, P. C., Fan, G. T., et al. (2014). In silico analysis on structure and DNA binding mode of AtNAC1, a NAC transcription factor from *Arabidopsis thaliana*. *J. Mol. Modeling*. 20 (3), 2117. doi: 10.1007/s00894-014-2117-8

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

